THERAPY WITHOUT THE THERAPIST? EXPERIMENTAL EVIDENCE ON AUTOMATING A RELATIONAL SERVICE*

Bennet Feld† Nicolas Roever‡

September 1, 2026

Abstract

A large share of those affected by mental illnesses go untreated. Psychotherapy, a first-line treatment, works through a collaboration between patient and therapist, and that relationship is why many stay away: therapy is expensive, and to be helped, one must disclose to another person and risk being judged. In a randomized controlled trial with 400 US adults with moderate-to-severe anxiety and depression, none in therapy and most of them lonely, we replace the human side of the collaboration with a voice-based AI delivering CBT. For this population, the treatment appears to survive the removal of the therapist: participants take up the AI-delivered care and their demand shifts away from the human provider, not away from care or social support. Assignment lowers anxiety and depression by roughly 0.8 standard deviations at four weeks, within the range reported for human-delivered CBT, moves the mechanisms CBT targets, and forms a working alliance matching published norms for human therapists. Demand for human therapy falls on a stated measure and, by 18 percentage points, in a behavioral choice, while demand for medication and reliance on friends and family are unchanged; among participants who prefer the AI, the stated appeal is the absence of another person's judgment. Treated participants also report more social contact, not less, and lower loneliness, with the contact gains concentrated among the loneliest. By removing the person, automation may expand access to relational services, lowering both their price and the social cost of seeking them.

Keywords: artificial intelligence, psychotherapy, relational services, automation, healthcare demand

JEL codes: I12, I18, O33, C93, D12, D83

*The trial was pre-registered on AsPredicted (#258,203, aspredicted.org/sv9ac7.pdf) and approved by the ethics committee of the University of Cologne. We thank Nava Ashraf, Oriana Bandiera, Felix Chopra, Julian Herpertz, Thomas Messer, Margot Paul, Marten Ritterath, Chris Roth, Matan Rubin, Christiane Szerman, Sebastian Tonke, Omry Yoresh, the Applied-Micro Seminar at the University of Cologne, the LSE Public Economics and Management groups, and the AI and Healthcare Meeting at the NBER for valuable comments; Lukas Wolf and Chris Aeberli for discussions about the application; and Frauke Gerster, Malte van der Heiden, and Moritz Lakenbrink for excellent research assistance. B.F. acknowledges funding from the Economic and Social Research Council and N.R. from the German Academic Scholarship Foundation; participant recruitment was funded in part by the owner of the application used in the study, and by C-SEB at the University of Cologne. A research agreement between the authors and the company guarantees the research team full control over the study's design, data, and analysis, and the freedom to publish the findings, whatever they are; the company had no right of review over this paper. The intervention used in this study was a previous version of the application. This version is distinct from, and does not reflect, the current commercial product of the company, which has since undergone substantial changes to its features, underlying models, and user experience. Accordingly, no finding, characterization, or conclusion reported herein is generalizable to, or constitutes an evaluation of, the product currently available to users. This paper is a scientific publication and none of its contents constitute marketing material.

†Corresponding author. Department of Economics, London School of Economics and Political Science, 32 Lincoln's Inn Fields, London WC2A 3PH, United Kingdom. Telephone: +44 7788 802162. Email: b.feld@lse.ac.uk

‡University of Cologne. Email: nicolas.roever@wiso.uni-koeln.de

Total word count: 12884.

I Introduction

Mental illness is a private experience: the worry that keeps a person awake and the effort of appearing fine at work are invisible to others. And it is a common one: among the most prevalent health conditions worldwide, it affects more than one in five US adults in a given year (GBD 2021 Diseases and Injuries Collaborators, 2024; National Institute of Mental Health, 2024b). It is also treatable. Psychotherapy is a first-line treatment, and its effects are well documented for anxiety and depression, the two most common diagnoses (Cuijpers et al., 2023; Hofmann et al., 2012). Yet about half of US adults with a mental illness receive no care (National Institute of Mental Health, 2024b).

Psychotherapy works through a relationship: patient and therapist collaborate, and the value of the treatment is created between them. Even cognitive behavioral therapy (CBT), a common and highly structured form, is manualized but not scripted (Cuijpers et al., 2023; Hofmann et al., 2012). The therapist guides, and the intellectual work comes from the patient, who learns to catch their own distorted thoughts and weigh them against the evidence of their life. Much of that evidence is private to the patient: past successes, relationships, and the texture of a situation. It must be drawn out in conversation rather than supplied. Consistent with this, the alliance between patient and therapist is among the most robust predictors of whether treatment succeeds (Horvath et al., 2011).

The same relationship keeps people away from the treatment. To be helped, one must disclose to another person something one would rather hide and risk being judged for it. Part of the cost is structural: therapy is delivered one-on-one by a scarce clinician, at about $143 per session out of pocket in the United States (Zhu et al., 2024). But barriers attached to the relationship itself are cited just as often. Among people who perceive a need for care and do not seek it, about one-fifth cite structural reasons such as cost and access, while 97 percent cite at least one attitudinal or relational reason (Andrade et al., 2014; Mojtabai et al., 2011). The most common is the wish to handle the problem on one's own, given by 73 percent of US respondents; the most widely reported stigma barrier is the fear of disclosure, named by 32 percent (Clement et al., 2015). A structural model of therapy demand reaches the same conclusion: making therapy free barely raises take-up, because the binding cost is the disutility of treatment itself (Cronin et al., 2025).

This paper studies what happens when a relational service is automated: which parts of the service an AI reproduces and how demand for the human provider responds. In a four-week randomized

controlled trial, we recruit 400 US adults with moderate-to-severe anxiety and depression, none currently in therapy, and randomly assign half to a voice-based AI application that teaches CBT and half to a psychoeducation control. The sample lies inside the treatment gap, and we first document what keeps it there. The barrier is not a lack of perceived need: most participants have prior experience with therapy, and most believe that in-person therapy would be very effective for them personally. However, open-ended elicitations reveal the same two kinds of barrier the surveys above describe. Structural barriers (cost, time, and access to a therapist) are the familiar targets of supply-side policy. Relational barriers (discomfort with human delivery, fear of judgment, and difficulty finding a good fit) attach to the human delivery itself (Clement et al., 2015). Both are prominent in our sample: our participants are precisely the population for whom care without the monetary and relational costs of a human therapist could be attractive. For this population, removing the therapist lowers both kinds of barriers at once. We find that the treatment appears to survive the removal and that demand shifts away from the human provider, not away from care or social support.

The intervention is a mobile application that teaches the techniques of CBT through spoken conversation with an AI. As in human-delivered CBT, the skills are taught not through generic instruction but through tailored, open-ended dialogue, the delivery mode that makes talk therapy effective and labor-intensive (Beck, 2021; Craske, 2017). Earlier digital tools could not sustain a conversation of this kind, and engagement with them was low (Baumel et al., 2019). This application is built to sustain it: users speak with the AI in sessions of five to twenty minutes, and a multi-agent architecture (separate agents deliver techniques, track progress, and maintain a case conceptualization) supports the conversation. Each session therefore builds on the last across the four weeks, the way a course of therapy does. Whether these sessions resemble therapy is a question we test rather than assume.

We first ask whether the application reproduces a therapist's effects on the symptoms of anxiety and depression. Engagement is the precondition, and it holds: with no compensation tied to use beyond the first week, participants spend a mean of 3.14 hours with the AI over four weeks, about the dose of a monthly course of CBT, and sustain it themselves. Assignment reduces anxiety and depression, with effects that grow over the trial to roughly 0.8 standard deviations at four weeks, within the range reported for human-delivered CBT (Cuijpers et al., 2025; Hofmann et al., 2012). We

also document a reduction in the specific mechanisms targeted by CBT: negative automatic thoughts and negativity bias fall, and behavioral activation rises. The relationship itself can be measured: on the standard alliance instrument, participants' scores match published norms for human therapists on every dimension, including the emotional bond (Munder et al., 2010). What predicts who benefits most, though, is not the bond but agreement on the concrete tasks of therapy: the collaborative work of the relationship, rather than the felt human connection.

The AI also draws demand away from human therapists. Assignment lowers stated demand for human therapy by 0.46 standard deviations, and in a behavioral choice with real consequences, treated participants are 18 percentage points less likely to request information about a human-led online therapy program, from 66 to 48 percent ($p = 0.001$). The shift is specific to the therapist: demand for medication and reliance on emotional support from friends and family are unchanged.

Why does the average participant want less human therapy? The channel that distinguishes the AI is a shift in preference, and in participants' own words it turns on a single feature: the presence of another person. Those who prefer the AI cite the relief of that person's absence: less fear of judgment, lower social-evaluative threat, and freedom from embarrassment. Those who prefer a human cite the loss: genuine connection, empathy, and being understood. For a substantial share, the absence of a human is the appeal rather than a cost. These participants are not withdrawing from care but exchanging one provider for another, keeping the collaborative work of therapy while shedding its evaluation; for others, that same absence is what they would miss. A second channel is mechanical: as symptoms improve, the need for care falls. We cannot causally identify a mediation, but both mediators shift with treatment, and a descriptive decomposition following Gelbach (2016) apportions the demand effect across them: about 21 percent to the preference shift and 28 percent to symptom improvement, with the remainder unexplained by the measures we have. The fall in demand is thus not only people needing less care; a meaningful part is people who, having used the AI, would rather keep using it.

We then weigh this substitution against three questions: what it does to human connection in general, what it costs to deliver, and what health benefit it produces. Trading a human relationship for an AI could leave people more isolated (Andoh, 2025). However, we find the opposite. The fall in demand is confined to the professional: treated participants lean less on a therapist but no

less on their own network, reporting more frequent social contact and lower loneliness. And the social gains arrive where the concern points: in the markedly lonely sample, the increase in social contact concentrates among the loneliest participants. On cost, the application delivered 3.14 hours of care at a marginal cost near $6.29 (excluding development), against roughly $347 to $491 for the same hours from a human therapist. On benefit, mapping each participant's symptoms onto Global Burden of Disease disability weights, assignment averts 3.87 disability-days per participant over the four weeks, worth about $1,060 at a standard $100,000 valuation of a disability-adjusted life-year, two orders of magnitude above the cost of delivering it.

Four features of the design qualify these estimates; Section VIII takes up each in turn. The trial is unblinded against a psychoeducation control, so expectancy in either arm could affect the comparison; we benchmark the estimates against waitlist-controlled trials of human CBT, which embed the same comparison. Most outcomes are self-reported, though measured social-desirability and experimenter-demand concerns differ little across arms, the estimates survive controlling for them, and a behavioral choice guards the substitution result. The sample is selected on need and on openness to AI, so the estimates speak to the potential adopters rather than to all who need care—a selection we quantify against a survey matched to the US adult population on age, gender, and ethnicity. And the horizon is four weeks, with substitution measured while access to the AI is free, so the results speak to demand with the alternative in hand. Neither of these concerns, however, is likely to explain the results taken together: a general inclination to misreport improvement would move every margin alike, while the estimates move only where CBT aims and demand falls only for the provider the AI replaces.

Literature Contribution We contribute to three literatures. The first is the behavioral and organizational economics of human–AI collaboration. This work studies AI as a complement to a worker: it raises productivity in customer support, coding, and writing (Brynjolfsson et al., 2025; Noy and Zhang, 2023), enters medicine as diagnostic and decision support for the clinician (Abaluck et al., 2026; Agarwal et al., 2023; Goh et al., 2024; Mullainathan and Obermeyer, 2022), and asks when people trust an AI's advice and how its presentation shapes that trust (Conlon and Schwardmann, 2026; Dietvorst et al., 2018). The closest study pairs a worker with an AI on a one-shot innovation task and finds the pairing matches the productivity of a human team and reproduces some of its

social experience (Dell'Acqua et al., 2025). We move the AI to the other side of the exchange. It does not assist the professional; it takes the professional's place so that the consumer collaborates with the AI directly in a service that unfolds through repeated interaction and delivers its output, the skills of CBT, through the relationship itself. In our trial, this substitution holds in part. The collaboration produces symptom effects within the range reported for human CBT, moves the same mechanisms, and forms an alliance matching published norms for human therapists, at a marginal cost far below the price of the human service. What it does not replace is the human connection itself: participants who value that connection prefer the human provider and say so.

The second is the economics of mental health and the evaluation of digital and AI-delivered interventions, a literature developed largely outside economics (Li et al., 2023; Sohn et al., 2026; Zhao et al., 2025; Zhong et al., 2024) and, more recently, within it (e.g., Angelucci et al., 2026; Shreekumar and Vautrey, 2024; Woerner et al., 2025). This line of work asks whether such tools improve symptoms. A companion paper to our trial aimed at a clinical audience reports the two-week effects on anxiety and depression and examines the safety of such services (Röver et al., 2026). The present paper adds the four-week effects and, in particular, the mechanisms: AI-delivered CBT moves the processes the method targets (negative automatic thoughts, negativity bias, and behavioral activation) and forms a measurable therapeutic alliance. The pattern is consistent with the treatment operating as CBT is understood to operate, rather than through a generic lift in mood. Closest is contemporaneous work by Angelucci et al. (2026), who trace the longer-run effects of a broad digital mental-health platform in Mexico that bundles several functions, including an AI chatbot for urgent emotional support. Our study is complementary: they follow a full platform over a longer horizon; we study a single-purpose agent built to imitate human CBT closely, which lets us speak to the interaction itself, the mechanisms it moves, and the alliance it forms.

The third is the take-up of services that carry a social or stigma cost. Many services go underused not because they are scarce but because seeking them means being seen: the discomfort, shame, and fear of judgment in asking another person for help (Bursztyn and Jensen, 2017; Bursztyn et al., 2025; Shreekumar and Vautrey, 2024). Psychotherapy is a clear case: only a fraction of those who would benefit ever start, deterred by cost and time but also by stigma (Clement et al., 2015; Cronin et al., 2025; Mojtabai et al., 2011; Thornicroft et al., 2017). The same friction lowers take-up of welfare (Friedrichsen et al., 2018), HIV testing (Yu, 2023), and care among the socially excluded

(Ghosal et al., 2022). Where this literature intervenes, it works on the friction itself, correcting beliefs about how others judge or repairing self-image. We study a different lever: removing the other person while keeping the service. This lets us measure how demand for the human service responds when the social cost of being seen is taken out of seeking care. In our sample the response is specific: demand falls for human therapy, not for medication or lay support, and participants' own explanations point to the absence of a human evaluator. Whether the same removal works for services where the relationship is more of the product itself, as in mentoring or caseworking, is a question we leave open.

The remainder of the paper is organized as follows. Section II describes how therapy works when a human delivers it and why so few of those who would benefit receive it. Section III describes the AI's implementation of CBT. Section IV lays out the experimental design, and Section V describes the data. Section VI presents the therapeutic outcomes, and Section VII discusses substitution away from human providers and social contact. Section VIII addresses robustness and external validity and weighs the intervention's costs against its benefits. We conclude in Section IX.

II Context

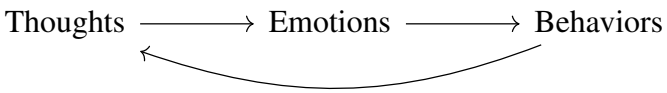
Anxiety and depression are the two most common mental illnesses in the United States: anxiety is the single most common class, and about 8 percent of adults have a major depressive episode in a given year (National Institute of Mental Health, 2024a). They are also among the most disabling conditions of any kind, accounting for a large share of years lived with disability worldwide and, at the individual level, predicting lower life satisfaction, worse physical health, and reduced employment and earnings (GBD 2021 Diseases and Injuries Collaborators, 2024; Ridley et al., 2020). The same conditions are among the most cost-effective to treat: every dollar invested in scaling up care for depression and anxiety returns an estimated two to three dollars in productivity alone (Chisholm et al., 2016). This section describes how that care works when a human delivers it and why so few of those who would benefit receive it.

II.A Cognitive Behavioral Therapy

Anxiety and depression respond well to both medication and psychotherapy (Cipriani et al., 2018; Cuijpers et al., 2023; Hofmann et al., 2012). Psychotherapy aims to teach skills for managing one's own well-being so its benefits can outlast the course of treatment (Hollon et al., 2006). Among the psychotherapies, cognitive behavioral therapy (CBT) has the largest evidence base for the two conditions we study (Cuijpers et al., 2023; Hofmann et al., 2012), and it is the approach our intervention teaches.

Clinical researchers divide what makes a therapy work into two kinds of ingredients: *specific factors*, the techniques particular to a given therapy, and *common factors*, the elements common to all of them, above all the relationship between patient and therapist (Cuijpers et al., 2019). This division organizes the section and section VI measures whether the AI moves both.

Specific Factors The specific factors of CBT are a set of cognitive and behavioral techniques practiced in session and through structured homework. They rest on the *cognitive model*, the framework through which a patient's case is conceptualized: a person's thoughts shape emotions, which drive behaviors, which in turn feed back into thoughts in a self-reinforcing cycle (Beck, 2021; Craske, 2017).



On this view, anxiety and depression are sustained less by events than by how people interpret them. Patients tend to generate *negative automatic thoughts*: rapid, involuntary appraisals that reflect systematic cognitive distortions rather than the situation at hand. After a poor examination, for instance, a person might conclude "I am a complete failure," overgeneralizing from a single setback, or expect every future attempt to end negatively, a *negativity bias*. CBT teaches patients to catch such thoughts and weigh them against the evidence, which lowers both their frequency and their emotional force. It works on behavior in parallel: through *behavioral activation*, patients deliberately re-engage in rewarding activities, which lifts mood and, through the same cycle, reshapes thought. We track all three as mechanisms below: negative automatic thoughts, negativity bias, and behavioral activation.

Common and Relational Factors Common factors are the elements common to most therapies: relational factors such as the *therapeutic alliance*, therapist characteristics such as empathy, and cultural factors (Wampold, 2015). Among them, the alliance, the quality of the working relationship between patient and therapist, stands out as one of the most robust predictors of treatment success in CBT as in psychotherapy generally (Cuijpers et al., 2019; Horvath et al., 2011). The alliance is usually separated into three dimensions, which we measure directly (Bordin, 1979): agreement on the goals of treatment, agreement on the tasks that pursue them, and the emotional bond between the two parties. The first two are the collaborative work of therapy; the third is the personal connection between patient and therapist.

Agreement on goals aligns the two on what treatment is for: the patient, as principal, and the therapist, as agent, settle on a shared aim. Agreement on tasks concerns the in-session work itself: which problems to focus on and which techniques to apply. CBT pursues both through collaboration rather than instruction. The therapist structures the session, selects the technique, and asks the questions; the patient supplies the material and does the thinking, treating their own beliefs as hypotheses to test rather than verdicts to accept. This division of labor is not incidental: much of the evidence a belief must be tested against is private to the patient, tacit, and imperfectly understood even by them, so it can only be drawn out through questioning, never supplied from outside. The aim is for patients to internalize the method and keep using it once sessions end, on their own and through homework. The bond, the third dimension, is the felt human connection between the two: mutual trust, liking, and attachment.

Both dimensions make therapy a relational service, and both are why it is commonly thought to require another person: the collaborative work of drawing out what is tacit and private and the bond of empathy and trust between two people, empathy being among therapy's more robust process predictors (Elliott et al., 2018). Relational services of this kind are regarded as among the hardest to hand to a machine (Garicano et al., 2026). In CBT, however, the techniques are the active ingredient, and the relationship is the collaborative platform through which they are delivered and practised (Craske, 2017). The treatment is delivered through a relationship without consisting of one. The patient therefore pays a relational price, the disclosure and the risk of judgment that another person's presence entails, for a service whose substance is guidance. How large that price is, and how often it keeps people from care altogether, is the subject of the next subsection.

II.B Access to Treatment

In 2022, about half of US adults with a mental illness received no treatment (National Institute of Mental Health, 2024b). Worldwide, only 16.5 percent of people with major depression received adequate care (Thornicroft et al., 2017). The reasons for this gap sort into three kinds: not perceiving a need for care; structural barriers to obtaining it, the familiar target of supply-side policy; and attitudinal or relational barriers to seeking it, such as stigma or doubt that therapy would help. The first is the largest single filter: 44.8 percent of US adults with a disorder who go untreated report that they did not think they needed care (Mojtabai et al., 2011). Among those who do perceive a need, structural barriers such as cost and access are cited by about a fifth of respondents, while nearly all name at least one attitudinal or relational reason (Andrade et al., 2014; Mojtabai et al., 2011). As Section V.C reported, trial participants overwhelmingly cite structural and relational reasons rather than a low perceived need.

Structural barriers The reliance on a skilled clinician that makes talk therapy work also makes it costly to deliver. Cost is the most cited structural barrier (Center for Behavioral Health Statistics and Quality, 2021): a session paid out of pocket runs about $143 in the United States (Zhu et al., 2024), roughly a week's groceries for the average household (U.S. Bureau of Labor Statistics, 2024), and the twelve to sixteen sessions recommended for CBT come to $1,700 to $2,300 (National Institute for Health and Care Excellence, 2011, 2022). Access is the second constraint: some 137 million Americans live in areas short of mental health providers (Kaiser Family Foundation, 2025). Yet these barriers explain little of the gap: in a structural model of therapy demand, eliminating the monetary cost of therapy has a negligible effect on take-up, suggesting that the binding disincentive is the disutility of therapy itself, not its price (Cronin et al., 2025).

Relational and attitudinal barriers Relational and attitudinal barriers are cited even more frequently than the structural ones. Among those who recognize a need but do not seek care, the single most common reason is a wish to handle the problem on their own, given by 72.6 percent in the United States and 63.8 percent across the WHO surveys (Andrade et al., 2014; Mojtabai et al., 2011). Stigma is close behind: across 144 studies, it stands among the main deterrents to help-seeking, with a median association of $d = -0.27$ (Clement et al., 2015). Much of that stigma

is the cost of being seen. The most widely reported stigma-related barrier is the fear of disclosure, of having something private become known, named by 32 percent of respondents (Clement et al., 2015).[1] The relational delivery of therapy thus appears on both sides of the ledger: it is how the treatment is traditionally delivered, and it is much of why it is costly to supply and hard to seek.

III Delivering Cognitive Behavioral Therapy with AI

Conversational AI has emerged as a potential solution to exactly this gap: cheap, private, and available on demand, it promises to lower both the structural and the relational barriers described above, and their adoption has been fast. People already use AI for emotional or mental-health support, both purpose-built apps and general-purpose assistants: surveys in the US and UK estimate that between 13 and 38 percent of adults have turned to an AI tool for such purposes (Cognitive FX, 2025; McBain et al., 2025; Mental Health UK, 2025).

We focus on the tools purpose-built for the role—mental-health companions colloqially know as AI therapists—rather than the general-purpose assistants people also turn to.[2] Many such tools have appeared over the past decade, from research groups and companies. Most are text-based chatbots that run on a single off-the-shelf language model, shaped by prompt engineering: written instructions that set the AI's persona and guide its replies. Woebot and Wysa are the best-known examples (Chang et al., 2024; Fitzpatrick et al., 2017; He et al., 2023).

The application we study belongs to the most recent generation. Instead of a single prompted model, it uses a fuller architecture that carries memory and structure from one session to the next and runs separate agents for specialized jobs like safety monitoring and case conceptualization. It is also spoken rather than typed. So in both what it does and how it does it, it comes closer to a human therapist than earlier tools.

The particular intervention used in this study is a commercially available application that was free of charge at the time of the study. It is voice-based, and users start sessions on demand, each

[1] Consistent with this, a national poll finds that 35 percent of US adults say they would view someone differently on learning that a person had a mental health condition (American Psychological Association, 2025).

[2] Terminology here is unsettled and shaped by regulation. As oversight of medical claims has tightened, providers increasingly describe these tools as "AI well-being companions" rather than therapists, and the medical literature uses careful formulations such as "an AI companion for the management of stress and emotional well-being using CBT principles". For readability we use the more colloquial "AI therapist".

lasting about 5–20 minutes, speaking with the AI as it guides them through five standard CBT building blocks: (1) a mood check-in, (2) collaborative agenda-setting, (3) a review of between-session practice, (4) the introduction and rehearsal of a CBT technique, and (5) a targeted homework assignment. Online Appendix Figure B.10 shows screenshots from the app.

Underneath sits a multi-agent architecture, in which specialized agents handle distinct roles: speech, technique delivery, case conceptualization, and patient history. This division of labor lets sessions coordinate and build on one another, so the intervention accumulates context across sessions rather than treating each in isolation.

Figure I shows excerpts from a first session, carried out by a member of the research team following an imagined case, with the version of the application used in the trial (see Online Appendix Section C for the full transcript). The session follows the structure of human-delivered CBT (Beck, 2021; Craske, 2017): a check-in, a collaboratively set agenda, an exploration of the chosen problem, the teaching of a technique, and assigned homework, omitting only the review of between-session practice, which has no place in a first session. Holding that structure together is what the multi-agent architecture is for. Asked about the same problem, a general-purpose assistant, built to be immediately helpful, would likely dispense advice; here, specialized agents keep the exchange inside a structured session, track the agenda across turns, and maintain a case conceptualization so that the AI follows the patient's thread, returns to it later (as in the closing turn), and introduces a technique when the moment is right rather than on demand.

The same session illustrates the two kinds of ingredient we distinguished earlier. The specific factors appear in the technique step, where the AI reframes the patient's way of coping through a metaphor and then helps them build a behavioral strategy. The common factors run through the whole exchange: its bond when the AI reflects the patient's feelings back, its goal dimension when the patient and AI settle together on what to work on, and its task dimension when they agree on the technique to pursue (Bordin, 1979). Throughout, the AI proceeds by open-ended questions rather than direct advice, so the patient reaches their own conclusions—the behavioral strategy is the user's suggestion, not the AI's (Braun et al., 2015).

[Figure 1 about here.]

IV Experimental Design

We ran a two-arm, five-wave randomized waitlist trial comparing the voice-based AI application to a psychoeducation control. Our experimental procedure was preregistered prior to data collection under AsPredicted #258,203; Figure II depicts the 4-week experimental timeline and the flow of participants across the five waves.

[Figure 2 about here.]

Recruitment and Screening (Wave 1 and Phone Screener) After clicking on our recruitment advertisement on Facebook and Instagram (see Online Appendix Figure B.1), individuals were redirected to a study website to complete a brief baseline questionnaire. Meta's ad-delivery algorithm chose who saw the advertisement. It learned from whether a viewer ultimately proved eligible, not from the eligibility criteria themselves, which we never disclosed, and so converged on the observable profile most predictive of engagement and eligibility.

Eligibility was determined primarily by participants' scores on the Generalized Anxiety Disorder-7 (GAD-7) scale, a widely used instrument for assessing anxiety symptoms (Spitzer et al., 2006). The study included individuals with GAD-7 scores between 13 and 20, ensuring that participants exhibited at least moderate levels of anxiety symptoms.[3] Additional pre-registered exclusion criteria were the following: participants had 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 excluding Nevada and Illinois, (5) not currently be 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. Participants meeting these criteria then completed an SMS-based two-step verification to screen out fraudulent responses and scheduled a phone call with a member of the research team.

The phone call followed a standardized script. During the call, operators explained the trial

[3]Scores of 21 were excluded as well, as this response pattern (selecting the highest option on every item) likely indicates low effort or inattention. Online Appendix Figure B.2 shows that the GAD-7 score criterion was the primary binding eligibility constraint.

procedure and confirmed that participants had sufficient time to take part. Operators also checked that participants did not have suicidal ideation and did not reside in Nevada or Illinois, which have introduced legislation limiting the provision of mental health apps. Those who met these criteria were invited to the Wave 2 survey.

Treatment Assignment and Conditions (Wave 2) As pre-registered, $N = 400$ participants were invited to complete Wave 2, which collected additional baseline variables not used for screening and then randomized participants in a 1:1 ratio to the treatment or control condition. Participants in the treatment arm received access to the voice-based AI application described in Section III. The control group received access to a dedicated web app containing a text-based summary of WHO material on mental health and stress management; the techniques it outlines are drawn from CBT and Acceptance and Commitment Therapy.[4] Assignment was not blinded, and the control group was told that the web app served as a substitute during the waitlist period. In both conditions, 98% of participants signed up for their assigned application (Figure II).

Outcome Collection (Waves 3, 4 and 5) Outcome measures were collected at follow-ups one, two, and four weeks after randomization. Every wave measured the main mental health symptoms (anxiety and depression symptoms) as well as core mechanisms. To limit participant burden in any single wave, the remaining survey measures were distributed across waves.

Payments Participants were told that they would receive up to $50 for completing all study waves, along with entry into a lottery to win an iPhone. The compensation schedule, communicated at the end of Wave 1, increased across waves: $5 for completing Wave 2 and signing up for the assigned application, and $10, $15, and $20 for completing Waves 3, 4, and 5. To ensure a minimum exposure to the intervention, the treatment group was additionally offered a $5 bonus for a first session of at least ten minutes; this bonus was ultimately paid to all participants, in both arms and regardless of completion, so no one was disadvantaged by it. To reduce attrition, we introduced two unannounced incentives: a surprise Thanksgiving lottery for an iPhone, announced between Waves 1 and 2, and a $10 gift card to a random subset of ten participants between Waves 4 and 5. Compensation tied to the

[4] The full app is available at: `https://github.com/NicolasRoever/web_app_mh`

intervention itself was thus confined to the first week and to initial take-up. All other compensation was tied to survey completion, so usage after the first week and session reflects participants' own decisions rather than financial incentives, a distinction that matters because engagement in trials of mental-health apps typically exceeds engagement in real-world use (Baumel et al., 2019).

Safety To protect the well-being of a vulnerable study population, we implemented a safety protocol. In summary, phone-based screening excluded participants with past-year suicidal ideation. During app onboarding, participants received transparent information about the intended role of the app as a supplementary, not substitutive, resource for individuals experiencing serious mental health difficulties. Both applications embedded direct links to professional support resources. The AI application included an automated safety mechanism that detected acute distress indicators through natural language processing; when crisis language or severe distress markers were identified, the system immediately ended the session and redirected users to appropriate professional services, including the 988 Suicide and Crisis Lifeline (Weber et al., 2026). Safety outcomes are reported and discussed in Röver et al. (2026).

V Data

V.A Outcome Measures

We collected outcomes at baseline and three follow-ups: one, two, and four weeks after randomization (Waves 3, 4, and 5). Each wave measured our primary symptom outcomes together with the core CBT mechanisms. The remaining survey measures were distributed across waves to limit participant burden in any single wave. Continuous outcomes are standardized to the baseline distribution so that estimated effects are expressed in standard deviations. All outcomes were pre-registered under AsPredicted #258,203 unless noted otherwise below; Online Appendix Table E.2 maps every pre-registered outcome to its place in the paper and Online Appendix Table E.1 documents the deviations from the pre-registration.[5]

[5]The companion paper, aimed at a clinical audience, reports the pre-registered primary symptom outcome (the two-week GAD-7 change) together with the two-week PHQ-8 and the trial's safety endpoints (Röver et al., 2026). The remaining pre-registered primary outcomes (detailed engagement and user satisfaction), the one- and four-week trajectories, the mechanisms and working alliance, the demand and social outcomes, the cost–benefit analysis, and all

Mental health outcomes We measure anxiety symptoms using the GAD-7, a pre-registered primary outcome (Spitzer et al., 2006). Alongside it we report depression symptoms, measured with the PHQ-8 (Kroenke et al., 2009) and pre-registered as a secondary outcome. Both are elicited at every follow-up. From the two scales, we also form a pre-registered composite mental-health index, the average of the separately standardized GAD-7 and PHQ-8 scores. Each item asks how often, over the past two weeks, the respondent has been bothered by a given symptom, rated on a four-point frequency scale from "not at all" (0) to "nearly every day" (3).

Usage and engagement Engagement with the intervention is itself a pre-registered primary outcome. Because compensation beyond the first session was tied to survey completion rather than to use of the application, observed usage reflects genuine engagement rather than financial inducement. From the application's logs we construct objective measures of engagement—the number of sessions, exchanges, and active days, and the cumulative conversation time with the AI (Figure Va)—and the analogous visit and time measures for the control web app (Online Appendix Figures B.6a and B.6b). Online Appendix Table A.8 reports each measure cumulated over the first week, the first two weeks, and the full trial. For the treatment arm we additionally record a complete transcript of every session (a representative example appears in Online Appendix Section C); these capture the content of each interaction and provide raw material for further analysis of how users engage with the intervention.

Mechanisms CBT is thought to work through several mechanisms, which we group, following the distinction in Section II, into common and specific factors. The first common factor we measure is the working alliance between the user and the application with the short-form Working Alliance Inventory (WAI-SR) at Wave 4 (Hatcher and Gillaspy, 2006). Because the instrument was written to capture the bond with a human therapist, we adapt each item to refer to the application rather than a clinician (e.g., "[the application] and I agree on what is important for me to work on") so that the scale measures the user's alliance with the AI. The working alliance is exploratory. The second common factor, engagement, is a pre-registered primary outcome and is described above.

For specific factors we measure the frequency of negative automatic thoughts (ATQ-8; Netemeyer et al., 2002), a validated instrument in wide use in clinical research, together with negativity bias

exploratory analyses appear only here; Online Appendix Table E.2 maps every pre-registered outcome to where it is reported.

and behavioral activation, each measured with two items asking how often, over the past two weeks, the relevant thought or behavior occurred. The negativity-bias items ask how often the respondent immediately expected an unclear or unexpected situation to turn out badly, and how often their attention was drawn to threats or negative cues rather than neutral or positive information. The behavioral-activation items ask how often they did something they genuinely enjoyed or found meaningful, and how often they put off or skipped activities because they felt low (reverse-scored). All three were pre-registered as secondary outcomes. For every multi-item construct, we average the items' scores and standardize against the baseline distribution so that effects are expressed in standard deviations.

Demand To study how the AI reshapes demand for conventional care, we measure it in two ways, following the survey literature on eliciting beliefs and preferences (Haaland et al., 2023; Stantcheva, 2023). First, we elicit stated demand for therapy with a human therapist and for psychiatric medication, asking individuals how likely they are to want each within the next month. We also elicit beliefs about the effectiveness of human-led and AI-based treatment, and willingness to pay for human and AI support. These items are not incentivized, and we test directly for experimenter-demand effects in Wave 5 (Online Appendix Section F). Second, we complement the stated measures with a revealed one: whether participants click to sign up for further information about a teletherapy provider, a real action rather than a hypothetical answer. The stated and revealed demand measures, the effectiveness beliefs, and the willingness-to-pay items were all pre-registered as secondary outcomes.

Perceptions of and Preferences for Artificial Intelligence We capture how participants evaluate the intervention along two dimensions. First, we collect a pre-registered battery of user-satisfaction measures: a self-reported satisfaction rating, willingness to pay for continued access to the application, and willingness to recommend it to a friend or family member. Second, we measure how participants in both groups evaluate AI relative to human care—their preference between AI and human psychological support and the reasons behind it, elicited in two steps following Haaland et al. (2025). In an open-ended item, participants explain in their own words why they prefer one modality over the other; we sort these responses into themes with BERTopic, which clusters dense

text embeddings, and refine the clusters into a coding scheme through researcher review (Online Appendix Section D). In a later wave we re-elicit the reasons as a closed multiple-choice item built from those themes, letting the categories emerge from participants' own language before we measure their prevalence at scale. The satisfaction measures were pre-registered as primary outcomes; the preference and reason items are exploratory.

Social outcomes We measure perceived loneliness with the three-item UCLA Loneliness Scale (Hughes et al., 2004), pre-registered as a secondary outcome and elicited at every follow-up. We additionally record the frequency of meaningful social contact and reliance on friends and family for emotional support, exploratory items we added mid-trial when high first-week engagement raised the concern of social displacement, a problem recently raised for AI tools more broadly (e.g., Cheng et al., 2026).

V.B Sample Descriptives

[Table 1 about here.]

Table I reports baseline characteristics by arm. The randomized sample is predominantly female (76% treatment, 81% control), with a mean age of about 39 years; employment is common but not universal (43% treatment, 50% control). Baseline symptom burden is high by construction, with mean GAD-7 scores in the severe range (15.9) and elevated PHQ-8 scores (14.1). Participants are also lonely: the mean score on the three-item UCLA Loneliness Scale is 7.6 in the treatment arm and 7.5 in the control arm of a possible 9, and 39% of the sample give the maximum score on all three items. Baseline covariates are balanced across arms for most variables; the exceptions are psychiatric-medication use, which is higher in the treatment arm ($p = 0.02$), and mean income, which is marginally lower in the treatment arm ($p = 0.06$). We thus conduct robustness checks for our main estimates by including these as additional covariates.

Attrition is low: only 2.5% (10 participants) by the 2-week follow-up and a similar 2.5% by the 4-week follow-up, which leaves little scope for attrition bias. Two checks, reported per effect in Online Appendix Section F, indicate that selective attrition is unlikely to account for the findings: a conservative baseline-carried-forward imputation for attriters leaves our estimates essentially

identical to the main specification, and Lee bounds cluster tightly around the point estimates (Lee, 2009, Online Appendix Figure F.1).

V.C Need for Therapy and Barriers in Sample

The sample is characterized by high clinical need alongside substantial demand for care. As Table I shows, participants enter the trial averaging in the severe range on the GAD-7 and with elevated depressive symptoms, and they are far from naive to the mental health care system: 75.5% have prior therapy experience and 71.5% have used psychiatric medication. Nor are they skeptical of the efficacy of conventional care: when asked to estimate the chance that "in-person therapy with a licensed psychologist" would improve their mental health, respondents give a mean of 57.8% (Online Appendix Figure B.3), and only a negligible share report no perceived need for therapy (Figure VIa).

Yet demand goes almost entirely unmet. None of our participants are in therapy at the start of the trial (by exclusion criterion), and only 3.8% take it up over the following four weeks. The gap between expressed demand and actual uptake points to access frictions.

We probe these frictions with open-ended elicitations (Haaland et al., 2025): in Wave 4 we ask participants why they are not currently in therapy and cluster the responses (Online Appendix Section D), and in Wave 5 we re-elicit them as a multiple-choice item built from those clusters (Figure III). The barriers fall into two groups. Structural barriers (cost, time, and access to a therapist) are cited most often, with affordability alone named by nearly half of respondents. Relational barriers (discomfort with human delivery, difficulty finding a good fit, and stigma) are mentioned at lower but still substantial rates and are harder to close through standard supply-side fixes such as subsidies or expanded provider networks. The same relational dimension surfaces when we ask participants with prior therapy experience what they valued and disliked about it (Online Appendix Figure B.4). What they valued is the relational core of in-person care (empathy, human presence, and validation), while their most common complaint, discomfort with human delivery, mirrors the relational barriers above.

[Figure 3 about here.]

Taken together, these patterns describe a population with substantial but largely unmet demand for care. Structural barriers are familiar targets for supply-side policy; relational barriers are not, and in our sample the same participants who report them also rate the human elements of in-person care most highly. Whether an on-demand AI tool can ease these frictions while preserving what people value about human therapy is an open question.

VI Therapeutic Outcomes

AI applications such as the one employed in this study promise to help improve mental well-being and build on CBT; if true, these apps could constitute a building block in addressing the need for mental healthcare and emotional support. First, we report intention-to-treat (ITT) effects of app assignment on anxiety and depression symptoms and benchmark them against published trials of human-delivered CBT. Second, we examine whether the AI tool adheres to the principles of CBT. Third, we assess the working alliance, a predictor of therapeutic success, against established norms from face-to-face therapy. Finally, we elicit how participants themselves perceive the AI relative to a human therapist.

VI.A Main Effects

We estimate intention-to-treat effects using our pre-registered specification:

$$Y_i = \beta_0 + \beta_1 \text{Treatment}_i + \beta_2 Y_{i,\text{baseline}} + \beta_3 X_i + \varepsilon_i \tag{1}$$

where Y_i denotes the outcome for participant i at the endline, Treatment_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 are heteroskedasticity-robust and clustered at the individual level. Standardization is done using the full baseline survey sample for GAD-7 ($N = 10,490$) and using the full randomized sample for PHQ-8 ($N = 400$).

Figure IV shows ITT effects of assignment to the AI app on anxiety (GAD-7) and depression (PHQ-8) symptoms across all follow-up waves expanding Röver et al. (2026). At 2 weeks, partic-

ipants assigned to the AI app score -3.52 points lower on the GAD-7 than in the control group ($p < 0.001$). The estimated reduction grows further by the 4-week follow-up, reaching -4.47 GAD-7 points, or -0.81 standard deviations ($p < 0.001$). These reductions are significant and large in absolute and clinical terms: the four-week reduction in the GAD-7 score exceeds the established minimal clinically important difference of 4 points (Toussaint et al., 2020), and the standardized estimates at both follow-up waves fall in the range that the clinical literature classifies as large (Hofmann et al., 2012). Effect sizes that reach this range after four weeks, when the benchmark trials tend to use longer courses of treatment, may partly reflect a motivated, self-selected sample and an always-available tool; Section VIII.A discusses the design features that could inflate the estimates and our checks against each. [6] These are intention-to-treat effects of assignment; the complier average causal effect, which rescales for imperfect take-up, is modestly larger and reported in Online Appendix Section F.2.

[Figure 4 about here.]

VI.B Mechanisms

Symptom changes on par with human CBT raise the question of mechanism: how does engagement with the AI produce these reductions? Little is known about how AI mental health and emotional support tools work (Li et al., 2023; Opel and Breakspear, 2026). We organize the evidence around the distinction from Section II. For the common factors, we measure the quality of the relationship between the AI and the user and the engagement that sustains it; for the specific factors, the established cognitive and behavioral mechanisms the method targets.

[Figure 5 about here.]

Engagement Unlike most forms of mental health support, participants fully choose and sustain their own use of the AI. Our design preserves this self-direction while ensuring that treated individuals familiarize themselves with the intervention. Beyond the first week and initial session, engagement reflects voluntary, self-administered use. On average, users invest a comparable amount of time to

[6] For a more extensive discussion of the primary effects on anxiety and depression symptoms, see our companion paper Röver et al. (2026).

a month of human CBT: over the four-week trial, mean cumulative conversation time with the AI is 3.14 hours, comparable to a usual monthly course of three to four 55-minute sessions, and 2.06 hours excluding the first week.

The dose arrives on a different schedule than human therapy. Instead of four weekly 55-minute sessions, participants use the tool in shorter, more frequent interactions: they open a voice session on a median of 9.0 days and talk for a median of 12.4 minutes on active days. Over the four weeks this amounts to a mean 15.4 sessions per participant (median 13). Thus, users engage with frequent, low-commitment contact, which would be hard to deliver with a human mental health professional.[7] Engagement is also skewed and heterogeneous (interquartile range 0.50–4.37 hours). Early drop-off is familiar from talk therapy, and similarly a mass of light users appears here; the contrast is in the upper tail, where about 10% of the treatment group spend more than eight hours with the AI over four weeks.

We report treatment heterogeneity by engagement as descriptive evidence in Online Appendix Section G, noting that variation in engagement is non-random and likely confounded through channels such as intervention-induced improvements in mental health. Treated participants with more conversational turns with the AI or a larger number of sessions see greater reductions in anxiety and depression symptoms.

Therapeutic Relationship We assess the alliance using the Working Alliance Inventory–Short Revised (WAI-SR; Hatcher and Gillaspy, 2006), the standard instrument for the three alliance dimensions introduced in Section II: agreement on goals, agreement on tasks, and the emotional bond. Participants assigned to the AI app report alliance levels that match the human benchmark of Munder et al. (2010) on every dimension: the mean total score is 3.75 (benchmark: 3.6), with subscale means of 3.69 for tasks (benchmark: 3.4), 3.81 for goals (benchmark: 3.6), and 3.75 for bond (benchmark: 3.7). The bond is the dimension where prior evidence predicts a gap: people value empathic communication from an AI less than from a human (Bernays et al., 2026; Kleinert et al., 2026; Rubin et al., 2025). We find no such shortfall on the bond subscale.

Higher alliance scores are associated with larger treatment effects, but the three dimensions

[7] In talking therapy, recent evidence suggests that weekly intensity (sessions per week) improves outcomes while the dose-response gradient measured by the total number of sessions or the length is flat (Cuijpers et al., 2013).

do not contribute equally. Online Appendix Table A.1 adds the standardized WAI-SR ratings to the right-hand side of our pre-registered ANCOVA while retaining the treatment indicator. A one standard deviation higher overall alliance rating is associated with a 0.18 standard deviation larger reduction in the mental-health index ($p < 0.01$). Entered one at a time, the goal and task subscales each predict a larger effect while the bond subscale does not; because the three subscales are highly correlated, these one-at-a-time results cannot attribute the association to any single dimension. Entered jointly, only the task subscale survives, while the goal and bond coefficients collapse to precise nulls. Agreement on the concrete tasks of therapy, rather than shared goals or the emotional bond, is thus what carries the predictive content of the alliance. Two limits bound this reading: the alliance is observed only in the treated arm, so these are associations among the treated rather than causal mediation, and prediction does not establish that task agreement produces the benefit.

Read together, the results point the same way. The AI reproduces CBT's symptom effects and mechanisms, and participants form an alliance at human norms. What predicts who benefits is not the felt bond but agreement on the tasks of therapy and baseline trust in the AI's accuracy, the credibility a patient would otherwise extend to a human therapist (Wampold, 2015). The pattern is consistent with the view that what CBT requires of the relationship is the collaboration it organizes rather than the empathic human who traditionally provides it (Zilcha-Mano, 2017), though our design cannot establish this directly.

Cognitive and Behavioral Mechanisms Figure Vc shows ITT effects on the canonical mechanisms targeted by CBT (Norbury et al., 2024). We find significant reductions in the frequency of negative automatic thoughts at every follow-up wave using an established construct from the literature (ATQ-8, Netemeyer et al. 2002), alongside significant improvements in a common cognitive bias—negativity bias. We also find evidence of behavioral activation: across all waves, participants assigned to the AI app report engaging in more enjoyable and meaningful activities. In short, engagement with the AI application activates the same mechanisms that drive symptom improvement in human CBT, including behavioral activation. [8]

[8]The effects on the mechanism outcomes are robust to the same battery of checks applied to the symptom outcomes: alternative specifications, baseline-carried-forward imputation, Lee bounds, and controls for social desirability and experimenter demand (Online Appendix Section F).

User perception Taken together, the AI matches human CBT on symptoms, mechanisms, and average dose, and the alliance users form with it is comparable to that reported with human therapists. Participants themselves also perceive that an AI can help their symptoms similarly to engagement with a human mental health professional. At the two-week follow-up, 85.1% of participants in the AI-app arm report that they perceive engaging with an AI to be at least as effective as human-delivered therapy. To rule out that this judgment is driven by inaccurate beliefs about human therapy, we exploit the fact that a majority of our sample has prior therapy experience: restricting the AI-app arm to participants with such experience, 87.2% report the AI as similarly or more effective than a human therapist.

Users also endorse the application on the pre-registered satisfaction battery. At the one-week follow-up, treated participants rate their satisfaction with the AI at a mean of 83.8 out of 100 and their willingness to recommend it to a friend or family member at 85.5 out of 100, the two closely correlated (Online Appendix Figure B.12). Stated willingness to pay for a 50-minute session with the AI averages $43.74 (Online Appendix Figure B.12a), well above its marginal cost of delivery (Section VIII.B).

VII Substitution and Social Contact

The previous section showed that the AI's effects on symptoms, the mechanisms behind them, and the relationship users form with it are comparable to what trials of human CBT report, and that users themselves perceive the AI as similarly effective to human therapy. We next ask how the availability of the AI reshapes demand for human care and thus whether it can help narrow the gap between the need for and the supply of mental health care. Section VII.A estimates how assignment shifts demand for human therapy, medication, and lay support; Section VII.B explores the channels behind the shift, quantitatively and qualitatively; and Section VII.C asks whether the substitution extends beyond professional care into participants' own social relationships.

VII.A Demand for Traditional Mental Healthcare

Table II reports estimates of how assignment to the AI app shifts demand for human therapy, elicited on a 0–100 scale and rescaled to lie between 0 and 1. In these survey items, we explicitly instruct participants to assume that a human therapist would be available to them. Across stated-demand items, participants assigned to the AI app express significantly less interest in seeking human-delivered therapy than participants in the control group. We also include a behavioral item carrying real consequences, which is plausibly less affected by experimenter demand and social desirability bias (Bursztyn et al., 2025; De Quidt et al., 2018); participants chose whether to receive information by email about online, human-led therapy programs ("Interest online therapy"). The estimated shift is large: participants in the AI-app arm are 18 percentage points less likely to request the information than controls ($p \leq 0.001$). Demand for psychiatric medication and for emotional support from friends and family, in contrast, is statistically indistinguishable across arms (Table II, "Demand for other care"), indicating that access to the AI app displaces human therapy specifically rather than mental health support more broadly.[9]

[Table 2 about here.]

Beyond stated intent, assignment shifts related beliefs and valuations in the same direction: it lowers participants' optimism about whether human therapy would help them and their willingness to pay for a session with a human therapist and raises their rating of the AI relative to a human provider (Online Appendix Tables A.2 and A.3). Belief in medication effectiveness was not elicited, so we cannot report a parallel effect there.

VII.B Mechanisms for the Change in Demand for Human Therapy

The demand for human therapy falls, and the fall is specific: demand for medication and reliance on friends and family do not move (Table II). The question is what produces this pattern. Our design cannot answer it causally: we lack independent variation in the mediators, and entering them in a regression would require that they be unconfounded with demand conditional on treatment and

[9]The demand and online-interest estimates are robust to alternative specifications, Lee bounds, and controls for social desirability and experimenter demand (Online Appendix Section F).

covariates. We therefore discuss each mechanism in turn and then combine them in a Gelbach decomposition read as descriptive.

Preference Against a Human Interlocutor The channel specific to this intervention concerns who delivers the care rather than what is delivered: Section VI suggests that the treatment's effects and mechanisms survive the substitution. What changes is the interlocutor. We ask each participant at Wave 4 whether they would prefer a human therapist or the AI, have no preference, or want no therapy at all. A majority in the treatment arm prefer the AI (Figure VI), consistent with the large majority who rate it as at least as effective as a human therapist (Section VI.B), and a multinomial regression shows that assignment shifts these preferences toward the AI; that is, treatment moves the mediator itself (Online Appendix Table A.4).

Open-text responses show what the preference turns on, and it is the presence of another person. Those who prefer a human predominantly cite human connection, empathy and true understanding, and perceived competence; those who prefer the AI predominantly cite fear of judgment and embarrassment, availability, perceived competence, and social avoidance (Figure VIb). This contrasts with the conventional clinical framing, in which the absence of authentic human empathy is treated as a fundamental constraint of AI-delivered support (Opel and Breakspear, 2026). For some, that absence is instead the central appeal: the reduction in social-evaluative threat outweighs the loss of human connection. It is also the impediment Cronin et al. (2025) identified from the demand side: reluctance to use therapy persists even when its monetary and time costs are removed, leaving the disutility of the treatment itself as the binding cost.

[Figure 6 about here.]

Need for Therapy A second, more mechanical channel is reduced need: participants who feel better need care less, whether because they can sustain the improvement with the skills they learned or through continued use of the AI app. The channel's premise holds in our data: symptoms improve (Section VI). On its own, it has a further implication: improved mental health should reduce the need for all symptom-directed care, so demand for psychiatric medication should fall alongside demand for human therapy. The data show an asymmetry instead: stated demand for human therapy falls by about 13 percentage points, while demand for medication moves little (Table II). Both start near the

middle of the scale (control means between 0.5 and 0.63), so the asymmetry is unlikely to be a floor effect. Squaring the asymmetry with reduced need requires an additional assumption: therapy and medication are discrete choices with their own thresholds and elasticities, so the same improvement can move one and not the other. The asymmetry therefore does not fully rule out reduced need.

Other Mechanisms Behavioral activation, a further mechanism, would work against the fall in demand rather than explaining it. Angelucci et al. (2026) find that use of their digital mental health platform increased take-up of psychotherapy, which they attribute to a lower subjective cost of seeking help consistent with behavioral activation. The AI raises activation in our arm as well (Figure Vc). To the extent this lifts demand for human therapy, this mediator would attenuate rather than explain the reduction we observe.

A Non-Causal Decomposition We cannot causally separate the mechanisms. A causal mediation analysis would require sequential ignorability: conditional on treatment and controls, the mediators must be as good as randomly assigned with respect to demand, so that $\hat{\gamma}_m$ recovers a mediator's effect rather than that of a common cause. In our setting this asks that, among participants assigned the same way and with the same baseline symptoms and history, a person's Wave 5 symptoms and Wave 4 stated preference be unrelated to whatever else drives their Wave 5 demand. Neither is guaranteed. A life event could improve symptoms and independently lower demand; the decomposition would fold its effect into the symptom channel. A shift in social avoidance could move both stated preference and demand; the preference channel would absorb it.

We therefore read the Gelbach (2016) decomposition as a descriptive accounting exercise: it apportions the treatment effect across measured mediators and offers a suggestive look at how demand for an established treatment reallocates when a new alternative appears. We enter mental health and preferences as the two channels, which amounts to treating behavioral activation as part of mental well-being.[10] An omitted channel independent of the included mediators leaves their contributions intact and enlarges the residual, while one correlated with an included mediator

[10]Entered as a third mediator, behavioral activation absorbs essentially none of the treatment effect beyond mental health (Online Appendix Table A.6), so its omission does not change the decomposition below. This does not show the channel is absent: activation is bound up with symptom improvement, so the two share variance and cannot be cleanly separated. It shows only that activation carries no contribution to demand beyond what the symptom composite already absorbs.

contaminates that mediator's $\hat{\gamma}_m$.

On the participants with complete mediator data, we run a short regression of Wave 5 demand on treatment and our pre-registered controls and a long regression that adds the two mediators: a Wave 5 standardized composite of GAD-7 and PHQ-8, S_i, and the Wave 4 preference indicators P_i^k (reference: prefers a human):

$$D_i = \delta^s T_i + \mathbf{X}_i'\beta + u_i, \tag{2}$$

$$D_i = \delta^f T_i + \gamma_S S_i + \textstyle\sum_k \gamma_k P_i^k + \mathbf{X}_i'\beta + \varepsilon_i, \tag{3}$$

where D_i is stated demand for human therapy, T_i is assignment to the AI app, and $\mathbf{X}_i$ the pre-registered controls.[11]

The gap between the two treatment coefficients, $\hat{\delta}^s - \hat{\delta}^f$, is the part of the effect the mediators absorb, and it splits cleanly across them. For each channel m we regress the mediator on treatment and controls to recover $\hat{\pi}_m$, the treatment-induced shift in that mediator and multiply it by $\hat{\gamma}_m$, its partial association with demand in Equation (3). The product $\hat{\pi}_m\hat{\gamma}_m$ is that channel's contribution, and the contributions sum to the absorbed gap, $\hat{\delta}^s - \hat{\delta}^f = \hat{\pi}_S\hat{\gamma}_S + \sum_k \hat{\pi}_k\hat{\gamma}_k$.

Table III reports the contributions. The two channels together account for about half of the demand reduction: symptom improvement absorbs 28% of the 12 percentage-point reduction, and the preference shift a further 21%, with the rest left in the residual effect. Within the preference channel, the contribution comes from one margin: the rising share of participants who prefer the AI relative to a human therapist accounts for about 26% of the treatment effect on its own; movement into indifference offsets a small part of this, and outright refusal of therapy contributes almost nothing, netting to the channel's 21%. A suggestive reading is that the fall in demand for human therapy comes both from people who need less care and from people who, having used the AI, would rather keep using it.

[Table 3 about here.]

[11] Specifically: an indicator for prior therapy experience, the Wave 2 measure of personal efficacy beliefs about human therapy, the GAD-7 baseline score (Wave 1), and the PHQ-8 baseline score (Wave 2).

VII.C Social Crowding-In

Substituting an AI for a human therapist raises a concern that reaches beyond care: if the AI stands in for one human relationship, it may also crowd out the others. Concerns of this kind have been raised for AI companions generally (Andoh, 2025; Cheng et al., 2026), and it arose in our own trial, where engagement in the first week was high enough to make displacement plausible. We therefore added two exploratory items mid-trial, the frequency of meaningful social contact and reliance on friends and family for emotional support. Perceived loneliness, measured with the UCLA Loneliness Scale (Hughes et al., 2004), was pre-registered as a secondary outcome and elicited at every follow-up.

Across the three outcomes, the pattern is the opposite of displacement. Reliance on emotional support from friends and family is unchanged (Table II). Treated participants report more frequent meaningful social contact. Table IV reports the intention-to-treat effect on the reported frequency of meaningful social contact, seven ordered categories ranging from "about once a week or less" to "many times a day." The ordered-logit coefficient is a log odds ratio, so the estimate of 0.64 means assignment multiplies the odds of reporting a higher contact frequency by $\exp(0.64) \approx 1.9$, at every threshold of the scale. Treated participants also report consistently lower loneliness at every follow-up, by 0.70 standard deviations at four weeks (Table IV; Online Appendix Figure B.8). The fall in demand documented in Section VII.A is thus confined to the professional: treated participants lean less on a therapist but no less on their own network.[12]

Exploratory heterogeneity by baseline loneliness, measured before randomization, sharpens this picture (Table IV). For participants below the maximum of the baseline UCLA-3 scale, the ITT is indistinguishable from zero. For the 39 percent at the scale maximum, the interaction adds 1.06 on the log-odds scale ($p = 0.005$): combined with the main effect, assignment roughly quadruples the odds that these participants report more frequent contact. The continuous version of the interaction shows the same gradient.[13] The reduction in loneliness itself does not sharpen the same way: the continuous interaction carries the aligned sign at every wave but is small and insignificant, and the median-split interaction is essentially zero (columns (5) and (6); Online Appendix Table F.12 reports

[12]The effects on loneliness and the social-interaction items are robust to the same battery of checks applied to the symptom outcomes: alternative specifications, baseline-carried-forward imputation, Lee bounds, and controls for social desirability and experimenter demand (Online Appendix Section F).

[13]The loneliness interaction is robust to competing treatment interactions with baseline symptom severity, to self- and social-image controls entered both directly and interacted with treatment, and to dropping respondents who report pressure to adjust their answers (Online Appendix Table F.11); attrition is unrelated to the treatment-by-loneliness cell.

every wave). At our sample size we cannot distinguish this null from a modest gradient, so we read the loneliness relief as broadly uniform and the concentration as a fact about behavior rather than about symptoms.

A suggestive reading is that behavioral activation extends to social life. CBT helps patients re-engage in activities they find meaningful; much of what people find meaningful is social, and Section VI documents that activation rises under treatment. The heterogeneity is consistent with this reading: the most isolated have the most room to re-engage and, plausibly, the most reason to.[14]

[Table 4 about here.]

VIII Discussion

VIII.A Robustness and Validity

Four features of the study warrant further investigation: the design is unblinded, the outcomes are largely self-reported, the sample is selected, and the horizon is four weeks. We first show that the estimates are not artifacts of specification or attrition, then probe each of the four in turn; the full set of tests is reported in Online Appendix Section F.

Alternative specifications and attrition. We probe each reported causal effect by adding baseline covariates one at a time, including the income and medication measures that are mildly imbalanced across arms, and the estimates barely move (Online Appendix Section F). Attrition is low: only 2.5% (10 participants) drop out by the two-week follow-up, at which we measure the primary outcome. For outcomes with a pre-treatment baseline, carrying that value forward for attriters leaves the estimate close to the main one (the carry-forward rows of the per-effect tables in Online Appendix Section F); for the demand items, first elicited after randomization, Lee bounds serve the same purpose. Across outcomes the bounds remain near the point estimates (Online Appendix Figure F.1), so modelling choices and selective dropout are unlikely to account for the results.

[14] A split by whether a participant's Wave-3 open-text explanation mentions fear of judgment or social avoidance shows no significant differentials for either outcome, but the classification is measured after randomization, and less fear of judgment is part of what the treatment produces, so we do not read it as moderation (Online Appendix Table F.13).

We make no multiplicity adjustment across outcomes. The primary outcomes were pre-specified in the pre-registration (Online Appendix Section E), the remaining pre-registered outcomes are reported as registered, and items added after registration are flagged as exploratory where they appear and, where applicable, estimated with the specification pre-registered for the primary outcomes.

Expectancy and anticipation under an unblinded design. Participants in psychotherapy trials can rarely be blinded to receiving treatment, and expectation is among the common factors of every talk therapy, human or AI (Wampold, 2015). Expectancy that produces real improvement in the treated arm is therefore not a bias but an ingredient of the treatment effect typically referred to as the placebo component. The comparison is, however, distorted if assignment itself moves the control arm. It could: frustration at waiting, or the salience of symptoms under repeated measurement, may have worsened control outcomes and inflated the estimated effect. But the direction is not given: the same salience, and the psychoeducational material, which 98% of controls signed up for, may just as well have caused improvements in the control group. As a hedge, we benchmark our estimates against waitlist-controlled trials of human CBT, which embed a similar comparison.

Assignment awareness can also enter the demand estimates directly: control participants answered knowing the app was coming to them. The natural channel is anticipation: a control participant deferring human therapy until the app arrives reports less demand for it, which biases the substitution estimate toward zero. The control arm's stated demand for human therapy is flat between the one-week and four-week follow-ups (65 to 63 percent), so we find no evidence that anticipation builds as access approaches; deferral held constant from the outset would not show in this comparison, but it too would bias the estimate toward zero. If anything, then, the substitution estimate errs on the conservative side.

What remains is expectancy in reporting: participants describing improvements and demand effects that did not occur. The pattern of the effects speaks against this: demand falls for human therapy while medication, if anything, drifts down, and lay support does not move (Table II); the cognitive and behavioral mechanisms move as CBT aims; and the same pattern extends to margins the app does not directly teach, loneliness and social contact. A general expectation of improvement is unlikely to single out each of these margins specifically, though fully ruling it out would require a blinding. The next paragraphs address misreporting from other sources.

Social desirability and experimenter demand in self-reported outcomes. Self-reported mental health raises a further concern: that respondents understate poor mental health to protect their self-image, or that treated participants infer the hypothesis and report what they think we want. Either would bias our estimates only if it differed across arms after treatment. We find little sign that it does. Self-image and social-image concerns, elicited in Wave 5 following Bursztyn et al. (2025), do not differ detectably across arms (Mann–Whitney $p = 0.16$ and 0.13), and, more to the point, the estimates hold when we control for them and when we drop participants who report feeling pressured to adjust their answers (the self-image, social-image, and experimenter-demand rows of the per-effect tables in Online Appendix Section F). And the specificity of the effects documented above cuts against these concerns as well, since a wish to please or to protect one's self-image would no more single out those margins than expectation would. For the substitution result, the behavioral demand measure below offers a further guard.

Stated and revealed demand. The fall in stated demand could mislead in two ways. The first is cheap talk: survey answers cost nothing. We therefore also measure a behavioral choice: treated participants are 18 percentage points less likely to request information about an actual human-led therapy program, from 66 to 48 percent ($p < 0.001$). The stakes of that choice are modest—information, not enrollment—but it is costlier than a survey answer, and behavioral measures of this kind are less exposed to experimenter demand than survey scales (De Quidt et al., 2018). The second is misperception: participants may not know what human therapy involves. The shift, however, is not confined to the therapy-naive: the same preference for the AI holds among the majority who have been in therapy (Online Appendix Figure B.7a).

Sample selection and external validity. Our sample is recruited through advertising and screened on openness to using AI and on symptoms, so it is not a random draw of US adults. Beyond these screening criteria, it differs from the general population in that participants are markedly lonely at baseline and disproportionately cite discomfort with human delivery and fear of judgment among their barriers (Section V.C). The selection is visible in the data: in a survey sample we recruited on Prolific and matched to the US adult population on age, gender, and ethnicity, the average control-arm participant would sit at the 96th percentile of the anxiety distribution and at the 83rd

percentile of comfort disclosing to an AI; Online Appendix Table H.1 reports the full comparison.

Each of these restrictions on external validity has a counterpart in the question we ask. Openness to AI is a prerequisite for treatment in the trial as in any real deployment, making these the potential adopters. Elevated symptoms leave room for improvement, making this a group that stands to benefit from treatment. And the prevalence of relational barriers lets us ask what removing the interlocutor does to access for a population in which social avoidance is a binding constraint. Nor are these characteristics rare: among people who perceive a need for care and do not seek it, nearly all cite at least one attitudinal or relational reason (Andrade et al., 2014; Mojtabai et al., 2011), and 45% of respondents in the same survey say they would be open to using a specialized AI for emotional support (Online Appendix Section H). One selection the design cannot reframe is on the gains themselves: those who expect an AI to help them are likelier to enroll, and consistent with this, effects are larger among participants with high baseline belief in the correctness of AI advice (Online Appendix Table A.11), so the effects we estimate are plausibly larger than the same treatment would produce among all potential adopters—though the human-CBT trials we benchmark against enroll treatment-seeking volunteers subject to the same selection. The boundary remains: within this population the estimates are informative; for those who would not disclose to an AI at all, and for those whom the presence of another person does not deter, they are silent.

Horizon. The trial lasts four weeks, and the limit reads differently for the two sets of results. For symptoms, the trajectory within the window rules out one worry: effects grow across the three follow-ups rather than fade, so the improvement is unlikely to be a novelty wearing off. Whether it endures beyond the window is a separate question, and our design cannot answer it directly. The mechanisms give grounds for expecting that it might: human-delivered CBT is among the few treatments whose effects persist after treatment ends, the accepted explanation being that patients internalize its skills (Hollon et al., 2006), and Section VI shows the AI moves the same mechanisms. Angelucci et al. (2026) document six-month persistence for a broader digital mental-health platform. For demand, the limit is more binding: substitution is measured during free access to the AI, so it shows that participants with the alternative in hand choose it over the human provider, not that they would if free access ends. That condition is, increasingly, the one consumers face: conversational AI tools for emotional support are available free or at low cost, and their use is already widespread

(McBain et al., 2025; Mental Health UK, 2025). What the design does not speak to is whether demand for human therapy recovers once the AI is withdrawn.[15]

VIII.B Cost-Benefit Analysis

We weigh the intervention two ways: the cost of delivering the same care through a human therapist and the health benefit it produces, valued through averted disability.

Cost of care delivered Our first measure asks what it would cost to deliver the same amount of contact through a human therapist. Answering it requires the assumption that an hour of conversation with the AI does the work of an hour with a therapist. Our second measure does not. Over the four weeks, participants spent a mean of 3.14 hours in conversation with the AI. Running the models costs at most about $2 per hour, so the variable cost of that care is roughly $6.29 per participant.[16] A therapist's time we price two ways, to bracket the comparison: the average US out-of-pocket rate, $143 for a standard 50-to-55-minute session, as an upper bound, and the Medicaid reimbursement, $82.77 for a 45-minute session, as a lower bound (Zhu et al., 2024).[17] Applied to the same 3.14 hours, the identical course of contact delivered by a therapist would cost between $347 and $491 per participant or 55 to 78 times the cost of the AI. Both figures value only the inputs to care, however, and say nothing about what those hours achieve.

Our second measure values the output directly, following the back-of-the-envelope approach of Francois et al. (2026). The Global Burden of Disease study assigns each disease state a disability weight between 0 and 1: the fraction of full health lost while living in that state, so that a year spent at a weight of 0.396, the GBD value for moderate depression, counts as 0.396 years of healthy life lost (Salomon et al., 2015). We map each participant's GAD-7 and PHQ-8 score at every wave into the weight for mild, moderate, or severe anxiety and depression, combine the two conditions with the standard multiplicative comorbidity correction so that the overlapping burden is not double-counted, and estimate the treatment effect on this comorbidity-corrected weight with our pre-registered ANCOVA at each follow-up. Table V reports the estimates: assignment lowers the disability weight

[15] Some participants explain their preference for the AI through availability (Section VII.B).

[16] The $2-per-hour figure is an upper bound on current model-inference costs and is meant to be illustrative.

[17] The Medicaid rate (CPT code 90834) runs about 40% below the average cash-pay rate. Spread over their session lengths, the two prices are about $110 and $156 an hour.

at every wave, and the effect grows over the trial. To aggregate the effect over the trial, write $\hat{\delta}_t$ for the estimated reduction on follow-up day t. On each study day the average treated participant carries a lower disability weight than the average control participant; summing these daily reductions over the 28 days, interpolating linearly between waves (the trapezoidal rule, anchored at zero on the day of randomization), gives

$$D \approx \sum_{k=0}^{K-1} \tfrac{1}{2}\left(\hat{\delta}_{t_k} + \hat{\delta}_{t_{k+1}}\right)(t_{k+1} - t_k) = 3.87$$

disability-days averted per participant. We value a disability-adjusted life-year at $100,000, following Open Philanthropy (2021); the anchor sits within the $100,000–$150,000 per quality-adjusted life-year that Neumann et al. (2014) suggest for the United States, and near the year of GDP per capita that Chisholm et al. (2016) use. A disability-day is then worth $100,000/365, and the benefit per participant assigned is $3.87 \times \$100{,}000/365 \approx \$1{,}060$, two orders of magnitude above the cost of delivering the AI.

Two underlying assumptions guide the interpretation of the estimate, and they can cut in opposite directions: setting the treatment effect to zero after four weeks is conservative and understates the benefit whenever any effect persists, whereas mapping self-reported symptom bands onto clinical disability weights is a strong assumption that could bias the benefit in either direction.

[Table 5 about here.]

IX Conclusion

Psychotherapy works through a relationship, and that relationship is why many of those who need it stay away: to be helped, one must disclose to another person and risk being judged. This paper asked what happens when the person is removed from the treatment. We ran a four-week randomized trial with 400 US adults inside the treatment gap, most of them lonely and many citing the discomfort of facing another person among their barriers. For this population, the treatment appears to survive the removal of the therapist: a voice-based AI teaching CBT reduced anxiety and depression by roughly the amount reported for a course of human-delivered CBT, moved the cognitive and behavioral mechanisms the method targets, and formed a working alliance that matched published norms for

human therapists. Demand, meanwhile, shifted away from the human provider, in stated intent and in a behavioral choice alike, but not away from care, medication, or the support of friends and family.

What the AI reproduced was the collaborative work of the therapeutic relationship rather than its felt human bond, and it is the collaborative work that predicts who improves. In removing the person, it removed the one thing a human provider cannot: the presence of another to disclose to and be judged by. Participants' stated preferences turn on exactly this feature: those who prefer the AI cite the relief of that person's absence and those who prefer a human cite the loss of connection. The social outcomes, meanwhile, run opposite to the displacement worry. Treated participants report more social contact and lower loneliness, not less, and the gains in social contact arise chiefly among the loneliest participants, the very group the worry concerns.

Four boundaries mark what these findings can claim. The comparison is against an unblinded psychoeducation control awaiting app access, and the outcomes are largely self-reported; Section VIII.A probes the resulting concerns, benchmarks the estimates against waitlist-controlled trials of human CBT, and shows the specificity of the effects cuts against reporting artifacts. The sample is selected on symptoms and on openness to AI, so the estimates speak to the potential beneficiaries rather than to all who need care. Our claim that participants received a treatment akin to therapy rests on fidelity to established CBT mechanisms rather than an audit of session content. And the four-week horizon cannot establish persistence, though the mechanisms give grounds to expect it and six-month persistence has been documented for a related intervention (Angelucci et al., 2026). The substitution result, in particular, is measured with free access to the AI in hand—a condition a growing share of consumers now faces.

For the population it reaches, removing the person from the treatment expands access to mental healthcare. The barriers that keep people from psychotherapy are financial, but at least as often they attach to the person delivering it: the discomfort of disclosing to another and the risk of being judged. An AI lowers both at once: it delivers the same hours of contact at a small fraction of a human therapist's cost, and it leaves no other person to face. The estimated value of the health improvements exceeds the delivery cost by two orders of magnitude. Consistent with this, the large majority of a sample that had stood inside the treatment gap took up and sustained the treatment,

with many citing the absence of judgment as its appeal. Whether the same holds for relational services where the relationship is more of the product itself (teaching, caseworking, mentoring) is a question our setting leaves open. But the mechanism may be general: many services go underused because seeking them means being seen, from HIV testing to welfare take-up (Friedrichsen et al., 2018; Ghosal et al., 2022; Yu, 2023), and where the observer is the deterrent, removing the person offers a lever that price cannot reach.

BENNET FELD, DEPARTMENT OF ECONOMICS, LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE, UNITED KINGDOM

NICOLAS ROEVER, UNIVERSITY OF COLOGNE, GERMANY

References

Abaluck, Jason, Robert Pless, Nirmal Ravi, Anja Sautmann, and Aaron Schwartz, "Does LLM Assistance Improve Healthcare Delivery? An Evaluation Using On-Site Physicians and Laboratory Tests," *NBER Working Paper No. 34660*, 2026.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," *NBER Working Paper No. 31422*, 2023.

American Psychological Association, "Public Attitudes Toward Mental Health in the U.S.," Survey conducted by The Harris Poll in partnership with the American Psychological Association; fielded April 18–21, 2025 among 1,076 U.S. adults May 2025.

Andoh, Efua, "Many teens are turning to AI chatbots for friendship and emotional support," *Monitor on Psychology*, 56 (2025), 51.

Andrade, L. H., J. Alonso, Z. Mneimneh, J. E. Wells, A. Al-Hamzawi, G. Borges, E. Bromet, R. Bruffaerts, G. de Girolamo, R. de Graaf, S. Florescu, O. Gureje, H. R. Hinkov, C. Hu, Y. Huang, I. Hwang, R. Jin, E. G. Karam, V. Kovess-Masfety, D. Levinson, H. Matschinger, S. O'Neill, J. Posada-Villa, R. Sagar, N. A. Sampson, C. Sasu, D. J. Stein, T. Takeshima, M. C. Viana, M. Xavier, and R. C. Kessler, "Barriers to mental health treatment: results from the WHO World Mental Health surveys," *Psychological Medicine*, 44 (2014), 1303–1317.

Angelucci, Manuela, Raissa Fábregas, and Antonia Vazquez, "The Well-Being Effects of Digital Mental Health Care," IZA Discussion Paper 18538, IZA Institute of Labor Economics 2026.

Baumel, Amit, Frederick Muench, Shir Edan, and John M. Kane, "Objective User Engagement With Mental Health Apps: Systematic Search and Panel-Based Usage Analysis," *Journal of Medical Internet Research*, 21 (2019), e14567.

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

Bernays, Florence, Marco Henriques Pereira, and Jochen Menges, "How Human is AI? Examining the Impact of Emotional Prompts on Artificial and Human and Responsiveness," *arXiv preprint arXiv:2601.05104*, 2026.

Bordin, Edward S., "The generalizability of the psychoanalytic concept of the working alliance," *Psychotherapy: Theory, Research & Practice*, 16 (1979), 252–260.

Braun, Justin D., Daniel R. Strunk, Katherine E. Sasso, and Andrew A. Cooper, "Therapist use of Socratic questioning predicts session-to-session symptom change in cognitive therapy for depression," *Behaviour Research and Therapy*, 70 (2015), 32–37.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond, "Generative AI at Work," *The Quarterly Journal of Economics*, 140 (2025), 889–942.

Bursztyn, Leonardo and Robert Jensen, "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure," *Annual Review of Economics*, 9 (2017), 131–153.

_ , Ingar K Haaland, Nicolas Röver, and Christopher Roth, "The Social Desirability Atlas," Technical Report, National Bureau of Economic Research 2025.

Center for Behavioral Health Statistics and Quality, "Key Substance Use and Mental Health Indicators in the United States: Results from the 2020 National Survey on Drug Use and Health," Technical Report HHS Publication No. PEP21-07-01-003, NSDUH Series H-56, Substance Abuse and Mental Health Services Administration, Rockville, MD 2021.

Chang, Christel Lynne, Chaitali Sinha, Madhavi Roy, and John Chee Meng Wong, "AI-Led Mental Health Support (Wysa) for Health Care Workers During COVID-19: Service Evaluation," *JMIR Form Res*, 8 (2024), e51858.

Cheng, Myra, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky, "Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence," *Science*, 391 (2026), eaec8352.

Chisholm, Dan, Kim Sweeny, Peter Sheehan, Bruce Rasmussen, Filip Smit, Pim Cuijpers, and Shekhar Saxena, "Scaling-up Treatment of Depression and Anxiety: A Global Return on Investment Analysis," *The Lancet Psychiatry*, 3 (2016), 415–424.

Cipriani, Andrea, Toshi A. Furukawa, Georgia Salanti, Anna Chaimani, Lauren Z. Atkinson, Yusuke Ogawa, Stefan Leucht, Henricus G. Ruhe, Erick H. Turner, Julian P. T. Higgins, Matthias Egger, Nozomi Takeshima, Yu Hayasaka, Hissei Imai, Kiyomi Shinohara, Aran Tajika, John P. A. Ioannidis, and John R. Geddes, "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 (2018), 1357–1366.

Clement, S., O. Schauman, T. Graham, F. Maggioni, S. Evans-Lacko, N. Bezborodovs, C. Morgan, N. Rüsch, J. S. L. Brown, and G. Thornicroft, "What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies," *Psychological Medicine*, 45 (2015), 11–27.

Cognitive FX, "Mental Health AI Chatbot Survey," `https://www.cognitivefxusa.com/blog/mental-health-ai-chatbot-survey` 2025.

Conlon, John J. and Peter Schwardmann, "AI Sycophancy and Decisions," Technical Report, SSRN April 2026.

Craske, Michelle G., *Cognitive-Behavioral Therapy* Theories of Psychotherapy Series, 2nd ed., Washington, DC: American Psychological Association, 2017.

Cronin, Christopher J., Matthew P. Forsstrom, and Nicholas W. Papageorge, "What Good Are Treatment Effects Without Treatment? Mental Health and the Reluctance to Use Talk Therapy," *Review of Economic Studies*, 92 (2025), 1699–1737.

Cuijpers, Pim, Clara Miguel, Mathias Harrer, Constantin Yves Plessen, Marketa Ciharova, David Ebert, and Eirini Karyotaki, "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 (2023), 105–115.

_ , **Marcus Huibers, David Daniel Ebert, Sander L. Koole, and Gerhard Andersson**, "How much psychotherapy is needed to treat depression? A metaregression analysis," *Journal of Affective Disorders*, 149 (2013), 1–13.

_ , **Mathias Harrer, Clara Miguel, Marketa Ciharova, Davide Papola, Djordje Basic, Cristina Botella, Ioana Cristea, Nino de Ponti, Tara Donker, Ellen Driessen, Pamela Franco, Irene Gómez-Gómez, Jessica Hamblen, Noelia Jiménez-Orenga, Eirini Karyotaki, Aaron Keshen, Jake Linardon, Emma Motrico, Minoo Matbouriahi, Olga M. Panagiotopoulou, Rory A. Pfund, Constantin Yves Plessen, Heleen Riper, Paula P. Schnurr, Marit Sijbrandij, Marieke B. J. Toffolo, Lingyao Tong, Wouter van Ballegooijen, Els van der Ven, Annemieke van Straten, Yingying Wang, and Toshi A. Furukawa**, "Cognitive Behavior Therapy for Mental Disorders in Adults: A Unified Series of Meta-Analyses," *JAMA Psychiatry*, 82 (2025), 563–571.

_ , **Mirjam Reijnders, and Marcus J. H. Huibers**, "The Role of Common Factors in Psychotherapy Outcomes," *Annual Review of Clinical Psychology*, 15 (2019), 207–231.

Dell'Acqua, Fabrizio, Charles Ayoubi, Hila Lifshitz-Assaf, Raffaella Sadun, Ethan R. Mollick, Lilach Mollick, Yi Han, Jeff Goldman, Hari Nair, Stew Taub, and Karim R. Lakhani, "The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise," Working Paper 33641, National Bureau of Economic Research 2025.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Management Science*, 64 (2018), 1155–1170.

Elliott, Robert, Arthur C. Bohart, Jeanne C. Watson, and David Murphy, "Therapist Empathy and Client Outcome: An Updated Meta-analysis," *Psychotherapy*, 55 (2018), 399–410.

Fitzpatrick, Kathleen K., Alison Darcy, and Molly Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," *JMIR Mental Health*, 4 (2017), e19.

Francois, Patrick, Matt Lowe, and Ieda Matavelli, "Psychedelics and Well-Being: An Experiment in Brazil," Working Paper, University of British Columbia 2026.

Friedrichsen, Jana, Tobias König, and Renke Schmacker, "Social Image Concerns and Welfare Take-Up," *Journal of Public Economics*, 168 (2018), 174–192.

Garicano, Luis, Jin Li, and Yanhui Wu, *Messy Jobs: The Work That AI Cannot Reach*, Upriver Press, 2026.

GBD 2021 Diseases and Injuries Collaborators, "Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *The Lancet*, 403 (2024), 2133–2161.

Gelbach, Jonah B., "When Do Covariates Matter? And Which Ones, and How Much?," *Journal of Labor Economics*, 34 (2016), 509–543.

Ghosal, Sayantan, Smarajit Jana, Anandi Mani, Sandip Mitra, and Sanchari Roy, "Sex Workers, Stigma, and Self-Image: Evidence from Kolkata Brothels," *The Review of Economics and Statistics*, 104 (2022), 431–448.

Goh, Ethan, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen, "Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial," *JAMA Network Open*, 7 (2024), e2440969.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, "Designing information provision experiments," *Journal of economic literature*, 61 (2023), 3–40.

_ , _ , Stefanie Stantcheva, and Johannes Wohlfart, "Understanding economic behavior using open-ended survey data," *Journal of Economic Literature*, 63 (2025), 1244–1280.

Hatcher, Robert L. and J. A. Gillaspy, "Development and validation of a revised short version of the Working Alliance Inventory," *Psychotherapy Research*, 16 (2006), 12–25.

He, Yuhao, Li Yang, Chunlian Qian, Tong Li, Zhengyuan Su, Qiang Zhang, and Xiangqing Hou, "Conversational Agent Interventions for Mental Health Problems: Systematic Review and

Meta-analysis of Randomized Controlled Trials," *Journal of Medical Internet Research*, 25 (2023), e43862.

Hofmann, Stefan G., Anu Asnaani, Imke J. J. Vonk, Alice T. Sawyer, and Angela Fang, "The efficacy of cognitive behavioral therapy: A review of meta-analyses," *Cognitive Therapy and Research*, 36 (2012), 427–440.

Hollon, Steven D., Michael O. Stewart, and Daniel Strunk, "Enduring effects for cognitive behavior therapy in the treatment of depression and anxiety," *Annual Review of Psychology*, 57 (2006), 285–315.

Horvath, Adam O, AC Del Re, Christoph Flückiger, and Dianne Symonds, "Alliance in individual psychotherapy," *Psychotherapy*, 48 (2011), 9–16.

Hughes, Mary Elizabeth, Linda J Waite, Louise C Hawkley, and John T Cacioppo, "A short scale for measuring loneliness in large surveys: Results from two population-based studies," *Research on Aging*, 26 (2004), 655–672.

Kaiser Family Foundation, "Mental Health Care Health Professional Shortage Areas (HPSAs)," KFF State Health Facts, from HRSA Bureau of Health Workforce data 2025.

Kleinert, Tobias, Marie Waldschütz, Julian Blau, Markus Heinrichs, and Bastian Schiller, "AI outperforms humans in establishing interpersonal closeness in emotionally engaging interactions, but only when labelled as human," *Communications Psychology*, 4 (2026), 23.

Kroenke, Kurt, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, 114 (2009), 163–173.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76 (2009), 1071–1102.

Li, Han, Renwen Zhang, Yi-Chieh Lee, Robert E. Kraut, and David C. Mohr, "Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being," *NPJ Digital Medicine*, 6 (2023), 236.

McBain, Ryan K., Robert Bozick, Melissa Diliberti et al., "Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults," *JAMA Network Open*, 8 (2025), e2542281.

Mental Health UK, "Over One in Three Using AI Chatbots for Mental Health Support," https://mentalhealth-uk.org/blog/over-one-in-three-using-ai-chatbots-for-mental-health-support-as-charity-calls-for-urgent-safeguards/ 2025.

Mojtabai, Ramin, Mark Olfson, Nancy A. Sampson, Robert Jin, Benjamin Druss, Philip S. Wang, Kenneth B. Wells, Harold A. Pincus, and Ronald C. Kessler, "Barriers to mental health treatment: results from the National Comorbidity Survey Replication," *Psychological Medicine*, 41 (2011), 1751–1761.

Mullainathan, Sendhil and Ziad Obermeyer, "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care," *The Quarterly Journal of Economics*, 137 (2022), 679–727.

Munder, Thomas, Fabian Wilmers, Rainer Leonhart, Hans Wolfgang Linster, and Jürgen Barth, "Working Alliance Inventory-Short Revised (WAI-SR): Psychometric properties in outpatients and inpatients," *Clinical Psychology & Psychotherapy*, 17 (2010), 231–239.

National Institute for Health and Care Excellence, "Generalised Anxiety Disorder and Panic Disorder in Adults: Management," NICE Clinical Guideline CG113 2011.

_ , "Depression in Adults: Treatment and Management," NICE Guideline NG222 2022.

National Institute of Mental Health, "Major Depression," https://www.nimh.nih.gov/health/statistics/major-depression 2024.

_ , "Mental Illness," September 2024.

Netemeyer, Richard G., Donald A. Williamson, Scot Burton, Dipayan Biswas, Supriya Jindal, Stacy Landreth, Gregory Mills, and Sonya Primeaux, "Psychometric properties of shortened versions of the Automatic Thoughts Questionnaire," *Educational and Psychological Measurement*, 62 (2002), 111–129.

Neumann, Peter J., Joshua T. Cohen, and Milton C. Weinstein, "Updating Cost-Effectiveness—The Curious Resilience of the $50,000-per-QALY Threshold," *New England Journal of Medicine*, 371 (2014), 796–797.

Norbury, Agnes, Tobias U. Hauser, Stephen M. Fleming, Raymond J. Dolan, and Quentin J. M. Huys, "Different components of cognitive-behavioral therapy affect specific cognitive mechanisms," *Science Advances*, 10 (2024), eadk3222.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 381 (2023), 187–192.

Opel, Nils and Michael Breakspear, "Transforming mental health research and care through artificial intelligence," *Science*, 391 (2026), 249–258.

Open Philanthropy, "Technical Updates to Our Global Health and Wellbeing Cause Prioritization Framework," `https://www.openphilanthropy.org/research/technical-updates-to-our-global-health-and-wellbeing-cause-prioritization-framework/` 2021.

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

Ridley, Matthew, Gautam Rao, Frank Schilbach, and Vikram Patel, "Poverty, depression, and anxiety: Causal evidence and mechanisms," *Science*, 370 (2020), eaay0214.

Röver, Nicolas, Lukas Wolf, Valerie Forman-Hoffman, Patricia Areán, and Bennet Lucca Feld, "Short-term Evaluation of a Voice-Based Generative AI Emotional Support Companion: A Randomized Controlled Trial," *PsyArXiv preprint*, July 2026.

Rubin, Matan, Joanna Z. Li, Federico Zimmerman, Desmond C. Ong, Amit Goldenberg, and Anat Perry, "Comparing the value of perceived human versus AI-generated empathy," *Nature Human Behaviour*, 9 (2025), 2345–2359.

Salomon, Joshua A, Juanita A Haagsma, Adrian Davis, Charline Maertens de Noordhout, Suzanne Polinder, Arie H Havelaar, Alessandro Cassini, Brecht Devleesschauwer, Mirjam Kretzschmar, Niko Speybroeck, Christopher J L Murray, and Theo Vos, "Disability weights for the Global Burden of Disease 2013 study," *The Lancet Global Health*, 3 (2015), e712–e723.

Shreekumar, Advik and Pierre-Luc Vautrey, "Managing Emotions: The Effects of Online Mindfulness Meditation on Mental Health and Economic Behavior," 2024.

Sohn, Jun-Seok, Byeong-Gwan Ha, SoHyun Park, Jae-Jin Kim, Eojin Lee, Hyangkyeong Oh, San Lee, and Eunjoo Kim, "Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms," *npj Digital Medicine*, 9 (2026), 377.

Spitzer, Robert L., Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe, "A brief measure for assessing generalized anxiety disorder: the GAD-7," *Archives of Internal Medicine*, 166 (2006), 1092–1097.

Stantcheva, Stefanie, "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible," *Annual Review of Economics*, 15 (2023), 205–234.

Thornicroft, Graham, Somnath Chatterji, Sara Evans-Lacko, Michael Gruber, Nancy Sampson, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Laura Andrade, Guilherme Borges, Ronny Bruffaerts, Brendan Bunting, Jose Miguel Caldas de Almeida, Silvia Florescu, Giovanni de Girolamo, Oye Gureje, Josep Maria Haro, Yanling He, Hristo Hinkov, Elie Karam, Norito Kawakami, Sing Lee, Fernando Navarro-Mateu, Marina Piazza, Jose Posada-Villa, Yolanda Torres de Galvis, and Ronald C. Kessler, "Undertreatment of people with major depressive disorder in 21 countries," *The British Journal of Psychiatry*, 210 (2017), 119–124.

Toussaint, Anne, Paul Hüsing, Antje Gumz, Katja Wingenfeld, Martin Härter, Elisabeth Schramm, and Bernd Löwe, "Sensitivity to change and minimal clinically important difference of the 7-item Generalized Anxiety Disorder Questionnaire (GAD-7)," *Journal of Affective Disorders*, 265 (2020), 395–401.

U.S. Bureau of Labor Statistics, "Consumer Expenditures – 2023," Economic News Release USDL-24-1948 2024.

Wampold, Bruce E., "How important are the common factors in psychotherapy? An update," *World Psychiatry*, 14 (2015), 270–277.

Weber, Samantha, Dustin Klebe, Lukas Wolf, Christopher Aeberli, Stephanie Homan, Nikolas Psathakis, Andrea Casanova, José Garcia Macias, Jack Sykstus, Jie-Ming Li, Barbora Provaznikova, Judith Rohde, Laura Frühschütz, Charlotta Rühlmann, Tobias Welt, Tobias Kowatsch, Birgit Kleim, MULTICAST consortium, and Sebastian Olbrich, "Suicide- and crisis-risk detection using large language models in mental-health chatbots," *medRxiv*, 2026.

Woerner, Andrej, Giorgia Romagnoli, Birgit M Probst, Nina Bartmann, Jonathan N Cloughesy, and Jan Willem Lindemans, "Should individuals choose their own incentives? evidence from a mindfulness meditation intervention," *Management Science*, 71 (2025), 7056–7070.

Yu, Hang, "Social Stigma as a Barrier to HIV Testing: Evidence from a Randomized Experiment in Mozambique," *Journal of Development Economics*, 161 (2023), 103035.

Zhao, Yuqing, Wei Qian, Yaru Chen, Donghong Wu, Yujia Luo, Cong Gao, Kankan Wu, and Zhengkui Liu, "Effect of an AI agent trained on a large language model (LLM) as an intervention for depression and anxiety symptoms in young adults: A 28-day randomized controlled trial," *Applied Psychology: Health and Well-Being*, 17 (2025), e70067.

Zhong, Wenjun, Jianghua Luo, and Hong Zhang, "The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis," *Journal of Affective Disorders*, 356 (2024), 459–469.

Zhu, Jane M., Aine Huntington, Simon Haeder, Courtney Benjamin Wolk, and K. John McConnell, "Insurance acceptance and cash pay rates for psychotherapy in the US," *Health Affairs Scholar*, 2 (2024), qxae110.

Zilcha-Mano, Sigal, "Is the alliance really therapeutic? Revisiting this question in light of recent methodological advances," *American Psychologist*, 72 (2017), 311–325.

Table I: 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
UCLA-3 Loneliness (mean)	7.6	7.5	0.5
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 AI app and control arms. Female is an indicator equal to 1 for respondents identifying as female. Medication indicates whether a respondent has ever taken psychiatric medication, coded as 1. Employment equals 1 for individuals working full-time or part-time. UCLA-3 Loneliness is the three-item UCLA Loneliness Scale (Hughes et al., 2004), scored 3–9. p-values report two-sample t-tests for equality of means across arms. Also reported in Röver et al. (2026).

Table II: Estimated effects of assignment to the AI app on demand for human therapy, medication, and lay social support

	Demand for human therapy			*Demand for other care*		
	Stated Week 1	Stated Week 4	Online interest	Medication Week 1	Medication Week 4	Emotional support
Treatment (AI app)	-0.130***	-0.127***	-0.182***	-0.001	-0.062*	0.151
	(0.032)	(0.031)	(0.051)	(0.036)	(0.036)	(0.188)
Control mean	0.652	0.628	0.663	0.509	0.543	—
Estimator	OLS	OLS	OLS	OLS	OLS	Ord. logit
N	347	389	390	364	389	390

Notes: Intention-to-treat estimates of assignment to the AI app under a modification of the pre-specified baseline specification. Controls include PHQ-8, GAD-7, prior therapy efficacy beliefs, therapy history, and medication history; no contemporaneous mental-health control is included. Stated demand and medication items are elicited on a 0–100 scale and rescaled to 0–1, so coefficients read as shares (a coefficient of -0.13 is 13 percentage points). Online therapy interest is a binary choice estimated by a linear probability model. The frequency of lay emotional support is ordinal and estimated by ordered logit; its coefficient is a log odds ratio, so $\exp(0.151) = 1.16$ multiplies the odds of reporting a higher support frequency, with a confidence interval that includes no change. No control mean is shown for this column because the mean ordinal category is not comparable to the shares in the other columns. Robust HC3 standard errors are used. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table III: Gelbach decomposition of the estimated effect on demand for human therapy

Component	Contribution	SE	% of treatment effect
Explained by improved mental health symptoms	−0.035		28.3%
		(0.015)	
via Preferences (all 3 dummies)	−0.027		21.4%
		(0.010)	
of which: AI therapy preference	−0.032		25.7%
		(0.011)	
of which: no preference	+0.006		-5.1%
		(0.005)	
of which: prefers no therapy	−0.001		0.8%
		(0.005)	
Residual direct effect	−0.062		50.3%
		(0.032)	
Total treatment effect (= sum of above)	−0.124		100.0%
		(0.030)	

Notes: Gelbach (2016) decomposition of the intention-to-treat effect of assignment to the AI app on stated demand for human therapy (Wave 5, 0–1 scale). All models, including the no-mediator baseline in the bottom row, are estimated on the $N = 386$ participants with non-missing mediators. Controls: baseline PHQ-8, GAD-7, prior beliefs about the efficacy of human therapy, prior therapy experience, and medication history. The mental-health channel is a standardized composite of Wave-5 GAD-7 and PHQ-8; the preference channel comprises dummies for preferring an AI therapist, having no preference, and preferring no therapy, each relative to preferring a human therapist (omitted), with sub-rows reporting each dummy's contribution. Each contribution is $\hat{\pi}_k \hat{\gamma}_k$, where $\hat{\pi}_k$ is the OLS treatment effect on mediator k and $\hat{\gamma}_k$ its partial effect on demand in the full model; standard errors (in parentheses) use the Gelbach (2016) GMM delta-method estimator.

Table IV: Treatment effects on social contact and loneliness

	Frequency of Social Contact (Wave 5, ordered logit)			UCLA Loneliness (Wave 5, OLS)		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.64***	0.31	0.70***	-0.70***	-0.72***	-0.70***
	(0.18)	(0.23)	(0.19)	(0.10)	(0.13)	(0.10)
High Loneliness (Wave 2)		-1.42***			0.06	
		(0.29)			(0.16)	
Treatment × High Loneliness (Wave 2)		1.06***			0.04	
		(0.38)			(0.20)	
UCLA Loneliness (Wave 2, std.)			-0.67***			0.71***
			(0.14)			(0.06)
Treatment × UCLA Loneliness (Wave 2, std.)			0.41**			-0.12
			(0.18)			(0.10)
Baseline Controls	✓	✓	✓	✓	✓	✓
Obs	390	390	390	389	389	389

Notes: Columns (1)–(3) report ordered-logit estimates (log-odds) for the Wave-5 frequency of meaningful social contact (seven-point scale): the average intention-to-treat effect, and interactions of treatment with baseline loneliness elicited in Wave 2 before randomization — column (2) an above-median indicator, which at the sample median of 8 marks participants at the scale maximum of 9, and column (3) the standardized continuous score. Columns (1)–(3) control for baseline PHQ-8, baseline GAD-7, beliefs about the efficacy of human therapy, prior therapy experience, and medication history. Columns (4)–(6) report OLS estimates for the standardized UCLA-3 loneliness score at Wave 5 with the same structure, controlling for baseline loneliness, beliefs about the efficacy of human therapy, and prior therapy experience; Online Appendix Table F.12 reports the loneliness interactions at every wave. Model-based standard errors for columns (1)–(3), heteroskedasticity-robust (HC3) for columns (4)–(6), in parentheses. * p;<;0.10, ** p;<;0.05, *** p;<;0.01.

Table V: Disability weights averted by assignment to the AI app

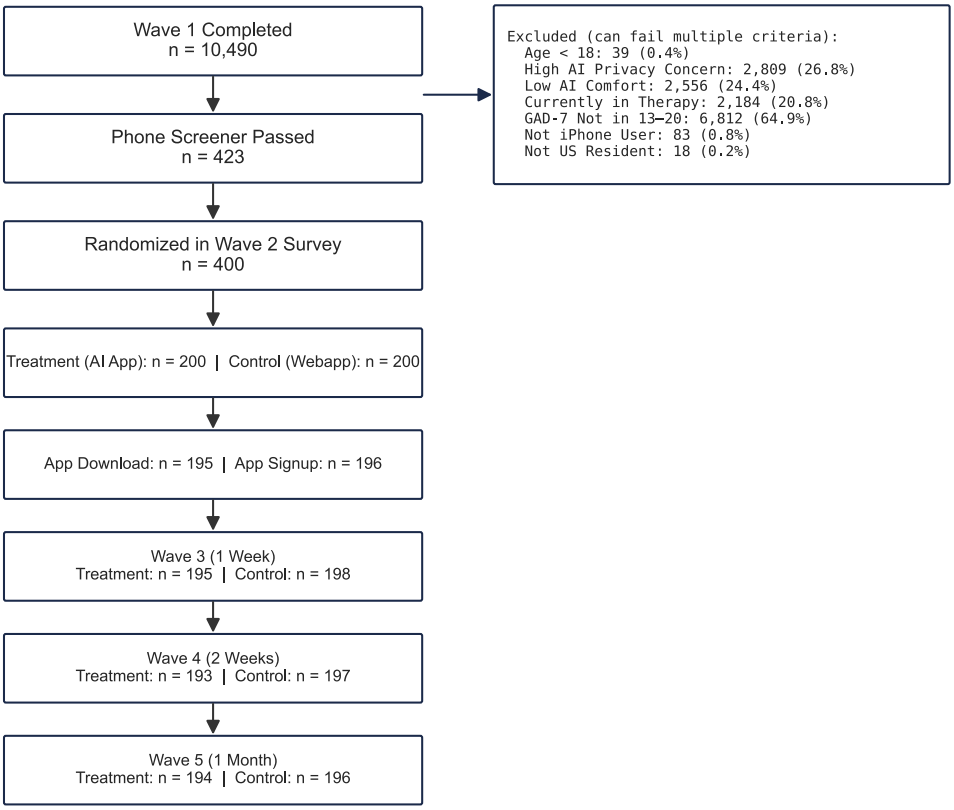
	Spec 1: pre-registered controls			*Spec 2: baseline disability weight*		
	Day 7	Day 14	Day 28	Day 7	Day 14	Day 28
Treatment (AI app)	-0.085***	-0.164***	-0.222***	-0.101***	-0.180***	-0.235***
	(0.022)	(0.025)	(0.025)	(0.023)	(0.026)	(0.026)
Disability averted ($-\hat{\beta}$)	0.085	0.164	0.222	0.101	0.180	0.235
Disability-days averted (0–28d)			3.87			4.24
N	393	390	390	393	390	390

Notes: Intention-to-treat estimates of assignment to the AI app on the comorbidity-corrected Global Burden of Disease disability weight (0–1), by follow-up day. Spec 1 uses the pre-registered controls (baseline GAD-7 and PHQ-8 entering separately); Spec 2 replaces them with the baseline combined disability weight. "Disability-days averted" integrates the per-wave effect over the four-week window (trapezoid, effect anchored at zero on the randomization day; assumed zero thereafter). Robust HC3 standard errors in parentheses. $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

1. Introduction and check-in	**AI:** [. . .] In this first touchpoint I want to get to know you better [. . .] what brought you here today? **User:** [. . .] I got rejected by a girl [. . .] asking myself if I'm able to really have long-term relationships. [. . .] I'm not really able to relax and calm down after work. **AI:** That sounds like a really draining mix—carrying the weight of that rejection while feeling like you can't fully recharge from work. *[. . .]*	**Alliance (bond)** The AI reflects the patient's feelings back, the beginnings of an affective connection.
2. Agenda setting	**AI:** What feels most pressing for you right now? **User:** I think relationships, definitely. *[. . .]*	**Alliance (goal)** Patient and AI agree on the session's focus together.
3. Exploration	**AI:** What feels hardest about it right now? **User:** Many people are already in established relationships, and the time's running out for me. [. . .] being lonely — it's loneliness. **AI:** How do you usually handle that loneliness? **User:** Seeking contact with anyone. [. . .] sometimes I'm sitting with a person thinking, what am I doing here — not with the people I really want. *[. . .]*	**Socratic questioning.** Open questions draw out the fear behind the rejection and loneliness, and how they cope with it.
4. Technique	**AI:** It sounds like you fill the space so you don't sit with uncomfortable feelings. [. . .] your energy is like a small garden, your attention the watering can — you pour water on every plant just to keep busy instead of tending the few that actually matter. [. . .] Which parts deserve a pour of energy? **User:** [. . .] really investing into being engaged with people when I meet them. **AI:** What small habit helps you arrive to those interactions with more energy? **User:** Taking a short deliberate break [. . .] setting an alarm ten minutes before I meet this person. *[. . .]*	**Cognitive reframing, built together.** The AI reframes how he copes with loneliness: reaching for any contact instead of tending the few that matter. **Behavioural strategy and alliance (task)** The patient generates a way to show up genuinely engaged, not drained.
5. Homework	**AI:** What would make following through realistic this week? **User:** Doing this in the morning, and when I plan my week on Sundays. [. . .] When I look on my calendar and realize I did it, I think this would help me. *[. . .]*	**Homework, set jointly.** The patient defines the plan *and* the sign that it is working.
6. Closing	**AI:** Before we wrap up, I'd like to circle back — you said relationships were what felt most pressing, and we haven't really unpacked that yet.	**Case conceptualisation & memory.** The AI tracks the agenda and reopens the goal it set earlier.

Figure I: Excerpts from a first session, annotated with the CBT principles at play. Full transcript in Section C.

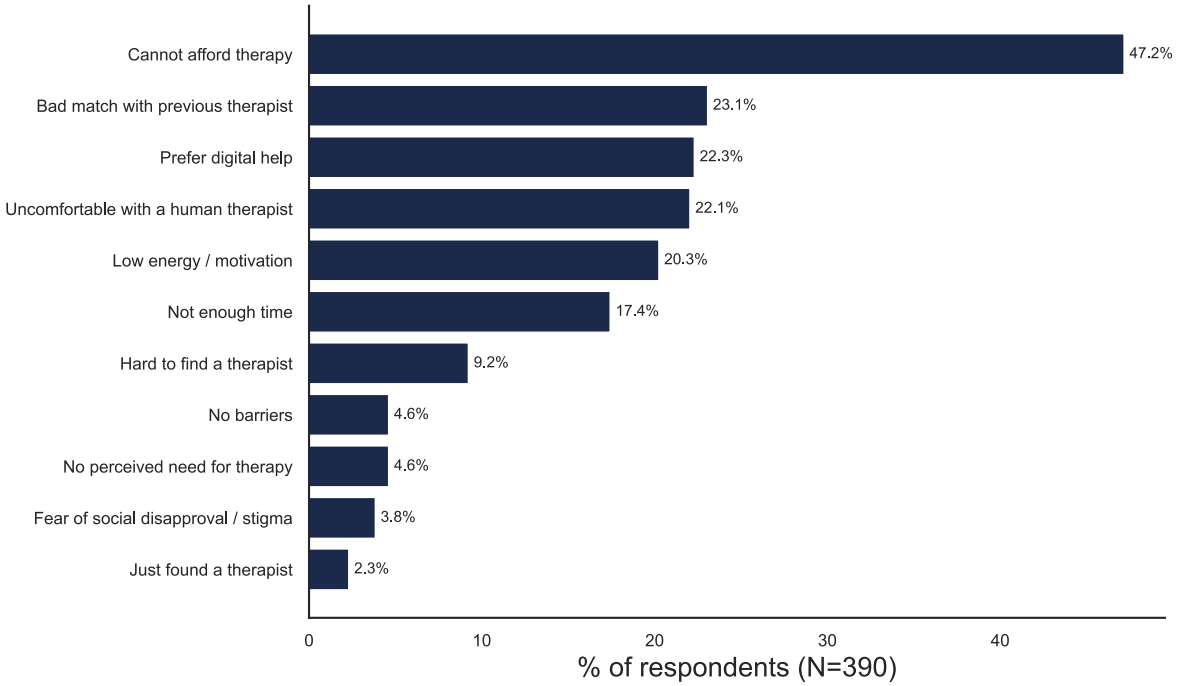
Figure II: Structure of the experiment



Note: Participant flow across the five survey waves. Of the 10,490 respondents completing the Wave 1 screening survey, 423 met all eligibility criteria and passed the phone screener and 400 were randomized in Wave 2, in equal numbers to the AI application and to the psychoeducation control. The box at the top right reports the exclusion criteria, which are not mutually exclusive: a respondent can fail more than one. The binding criterion is the GAD-7 range of 13–20, which excludes 6,812 respondents; screening on comfort with and privacy concerns about AI excludes a further 5,365, and current therapy 2,184. Follow-up waves are at one week (Wave 3), two weeks (Wave 4), and one month (Wave 5). Attrition is low and similar across arms, leaving between 193 and 198 participants per arm at every wave.

Alt text: Flow diagram of participant progression. Of 10,490 people who completed the screening survey, 423 were eligible and 400 were randomized, split evenly between the application and the control. A side box lists exclusion reasons.

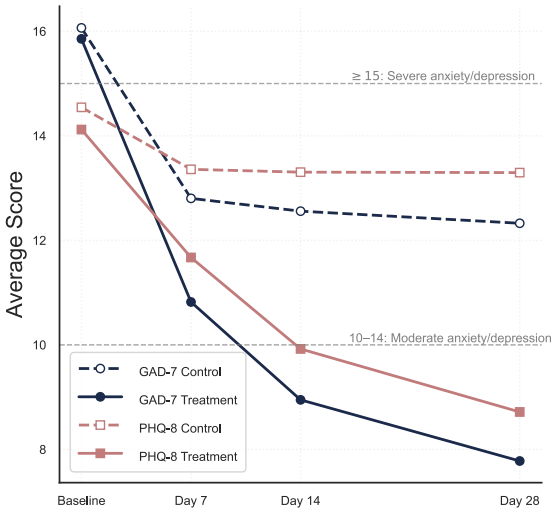
Figure III: Barriers to accessing therapy



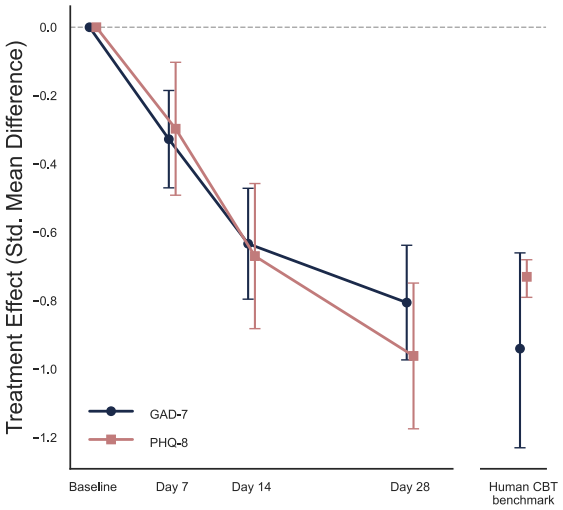
Note: This elicitation is done using a multiple-choice format in Wave 5. The categories were extracted from open-text responses in the previous Wave 4, using the procedure described in Online Appendix Section D.

Alt text: Horizontal bar chart of eleven reasons for not being in therapy. Cost is the longest bar followed by "bad match with previous therapist", preference for digital help, and discomfort with human therapist.

Figure IV: Average development of GAD-7 and PHQ-8 scores across arms and estimated intention-to-treat effects over time



(a) Average changes in PHQ-8 and GAD-7 trajectories.



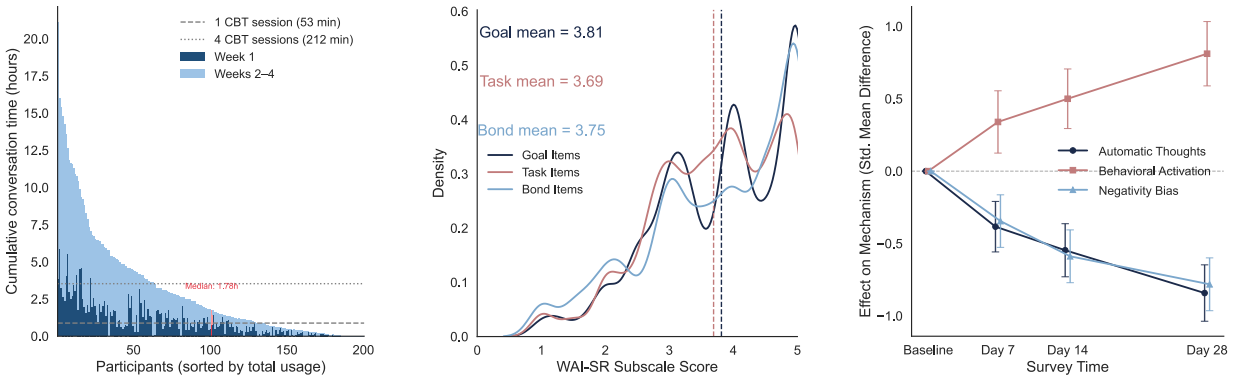
(b) Estimated intention-to-treat effects over time (standardized).

Note: Figure IVb reports standardized treatment effects at each wave with 95% confidence intervals estimated using our pre-specified ANCOVA model (see Online Appendix Section E). Two-week effects on anxiety (GAD-7) and depression (PHQ-8) from Röver et al. (2026). The human CBT benchmark is derived from RCTs broadly comparable to our trial conducted with human therapists from the *MetaPsy* database (see Online Appendix Figure B.5 for details).

Alt text: Panel (a). Line chart of average anxiety and depression scores at baseline and three follow-ups. Both treatment lines fall steeply and continuously; both control lines flatten after the first week.

Alt text: Panel (b). Estimated treatment effects on anxiety and depression at each follow-up, with confidence intervals. Both move steadily further below zero, ending near the human therapy benchmark plotted at the right.

Figure V: Estimated effects of assignment to the AI app on engagement, therapeutic alliance, and CBT mechanisms



(a) Engagement (hours).

(b) Therapeutic alliance (WAI-SR).

(c) ITT on cognitive and behavioral mechanisms.

Note: Left panel: cumulative conversation time with the AI application in hours over the four-week trial period; Online Appendix Table A.8 reports sessions, active days, and conversation hours cumulated over the first week, the first two weeks, and the full trial. Middle panel: distribution of Working Alliance Inventory (WAI-SR) goal, task, and bond subscale scores in the treatment arm at Wave 4. Right panel: estimated intention-to-treat effects of assignment to the AI app on the canonical CBT mechanisms estimated following the pre-specified ANCOVA (see Online Appendix Section E).

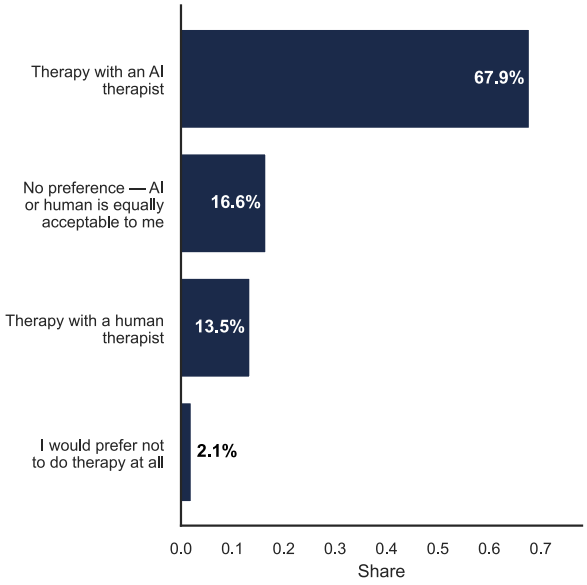
Alt text: Panel (a). Bar chart of total conversation time for each treated participant, sorted from highest to lowest, with the first week shaded separately. The distribution is strongly right-skewed.

Alt text: Panel (b). Three overlapping density curves, for the goal, task, and bond subscales of the working alliance measure. All three mass toward the top of the scale, with means close together.

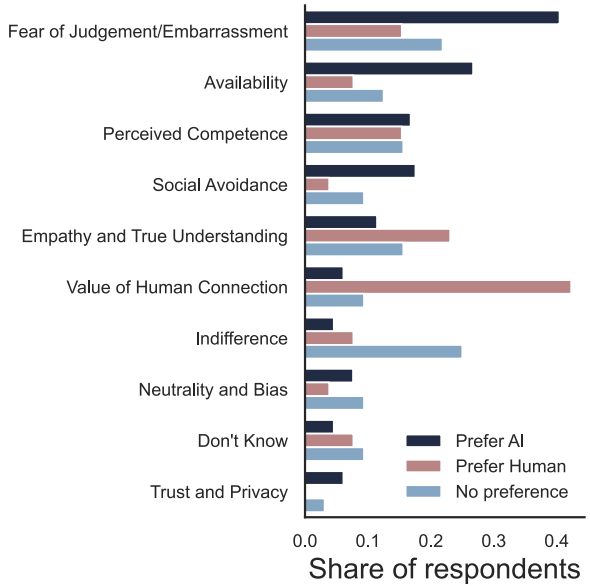
Alt text: Panel (c). Estimated treatment effects on three therapy mechanisms at each follow-up, with confidence intervals. Behavioral activation rises steadily above zero while automatic thoughts and negativity bias fall together below it.

Figure VI: Preferences for AI versus human therapy among participants assigned to the AI app

(a) Overall therapy preference (single-choice item)



(b) Reasons stated as relevant for therapy preference



Note: Panel (a) shows the distribution of responses among participants in the AI-app arm to a single-choice item on therapy modality preference, elicited at Wave 4: AI-delivered therapy, human-delivered therapy, no preference, or no current need for therapy. Panel (b) shows the share of participants whose open-ended Wave 3 explanations were classified into each of ten thematic categories using GPT-Luna 5.6. The bars are split according to respondents' Wave 4 answer to the survey question asking whether they would prefer therapy with an AI therapist, therapy with a human therapist, or had no preference. The graph excludes four respondents who indicated in Wave 4 that they would prefer not to receive therapy at all for better readability of the graph. Classification details are reported in Online Appendix Section D; the full taxonomy is given in Online Appendix Table A.5.

Alt text: Panel (a). Horizontal bar chart of therapy preference among treated participants. Preferring the application is by far the longest bar; preferring no therapy at all is the shortest.

Alt text: Panel (b). Grouped horizontal bar chart of reasons given for therapy preference, with separate bars for those preferring the application, a human therapist, or neither. Fear of judgment and valuing human connection are the longest.

Appendix:

THERAPY WITHOUT THE THERAPIST? EXPERIMENTAL EVIDENCE ON AUTOMATING A RELATIONAL SERVICE

Bennet Feld Nicolas Roever

August 28, 2026

Summary of the Appendix

- Section A collects the additional tables referenced in the main text.
- Section B collects the additional figures referenced in the main text.
- Section C contains a complete sample session transcript with the AI application.
- Section D describes our classification workflow for open-ended text data.
- Section E describes our statistical analysis plan and deviations from the pre-registration.
- Section F presents a uniform battery of robustness checks for every estimated effect.
- Section G reports engagement-based heterogeneity in estimated treatment effects.
- Section H compares the trial sample to a survey representative of US adults in gender, age, and ethnicity.
- Section I reports additional sample descriptives, including the sample selection process.
- Section J reports the cross-randomized information treatment.
- Section K contains the pre-registration document.
- Section L contains the full survey instructions.

A Additional Tables

Table A.1: Heterogeneity in Wave 5 Mental Health by WAI-SR (Wave 4)

	Pooled Subscales	Goal Subscale	Bond Subscale	Task Subscale	Total Score
Treatment ×					
WAI-SR Goal (Wave 4, std.)	0.06	-0.17**			
	(0.14)	(0.07)			
Treatment ×					
WAI-SR Bond (Wave 4, std.)	0.07		-0.11		
	(0.09)		(0.07)		
Treatment ×					
WAI-SR Task (Wave 4, std.)	-0.33***			-0.23***	
	(0.12)			(0.07)	
Treatment ×					
WAI-SR Total (Wave 4, std.)					-0.18***
					(0.07)
Treatment	-0.86***	-0.86***	-0.86***	-0.86***	-0.86***
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
ANCOVA Controls	✓	✓	✓	✓	✓
Obs	386	386	386	386	386
Adj. R^2	0.39	0.38	0.37	0.39	0.38

Notes: The dependent variable is the mental health composite score at Wave 5. Each column reports an OLS regression on the full sample (both arms), with the pre-registered ANCOVA controls: the Wave 1 mental health composite, prior therapy experience, and belief in the efficacy of human therapy (Wave 2). The WAI-SR is measured only for participants assigned to the AI app, so each Wave-4 alliance dimension is standardized within the treated arm and set to zero for the control group; its interaction with treatment therefore measures how the treatment effect scales with alliance, a treatment-intensity term. Column (1) interacts treatment with all three WAI-SR subscales (Goal, Bond, Task) jointly; columns (2)–(4) interact treatment with each subscale separately; column (5) uses the WAI-SR Total score. *Treatment* denotes assignment to the AI app, and its coefficient is the effect at average alliance. Standard errors (HC3, heteroskedasticity-robust) are reported in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table A.2: Estimated effects of assignment to the AI app on beliefs about treatment effectiveness

	Human belief (W3)	Human belief (W4)	Human belief (W5)	AI vs. human (W3, 1–5)
Treatment	-4.69**	-4.19*	-4.10	0.22**
	(2.16)	(2.28)	(2.53)	(0.10)
Baseline controls	✓	✓	✓	✓
Obs	393	390	390	392
Adj. R^2	0.31	0.24	0.20	0.05

Notes: Intention-to-treat estimates of assignment to the AI app. Columns (1)–(3) are ANCOVA regressions of the belief that a session with a human therapist would improve one's mental health (0–100 scale) at each follow-up. Column (4) regresses a 1–5 rating of AI relative to human therapy ("much less" to "much more effective", Wave 3) on assignment. All specifications use the pre-registered controls (baseline symptoms, prior therapy experience, baseline belief in human-therapy efficacy, and medication history). Belief in medication effectiveness was not elicited. Robust HC3 standard errors in parentheses. $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$.

Table A.3: Estimated effects of assignment to the AI app on willingness to pay for human therapy

	WTP human (W3)	WTP human (W4)	WTP human (W5)
Treatment	-3.47	-4.88	-6.90
	(5.32)	(4.94)	(4.81)
Baseline controls	✓	✓	✓
Obs	393	390	390
Adj. R^2	0.30	0.26	0.27

Notes: ANCOVA intention-to-treat estimates of assignment to the AI app on stated willingness to pay (USD) for a session with a human therapist at each follow-up, controlling for baseline willingness to pay, prior therapy experience, and baseline belief in human-therapy efficacy. Willingness to pay for the AI was elicited only in the treated arm and is reported descriptively in Figure B.12a. Robust HC3 standard errors in parentheses. $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$.

Table A.4: Effect of assignment to the AI app on stated therapy-modality preference (Wave 4)

	Spec. (2)		
	AI vs. Human	No pref. vs. Human	No therapy vs. Human
Treatment	0.58**	-0.31	0.55
	(0.29)	(0.34)	(0.81)
GAD-7 Score Wave 1 (std.)	-0.10	0.55	0.05
	(0.40)	(0.45)	(0.89)
PHQ-8 Score Wave 2 (std.)	0.10	-0.03	-0.60
	(0.16)	(0.18)	(0.51)
PHQ-8 (W2) baseline	Yes	Yes	Yes
N	390	390	390
Pseudo R^2	0.026	0.026	0.026
Average marginal effect of treatment assignment on outcome probability:			
P(Human therapy)		-0.042 (0.036)	
P(AI therapy)		+0.168*** (0.046)	
P(No preference)		-0.131*** (0.041)	
P(No therapy)		+0.005 (0.013)	

Multinomial logit, reference category: Therapy with a human therapist.

Robust (HC0) standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Notes: Multinomial logit of stated therapy-modality preference at Wave 4 on assignment to the AI app, controlling for baseline GAD-7 (Wave 1) and PHQ-8 (Wave 2); the reference category is preferring a human therapist. The upper panel reports coefficients for each alternative relative to the reference, interpretable as log odds ratios ($\exp(\beta)$ multiplies the odds of the alternative relative to preferring a human therapist); the lower panel reports the average marginal effect of assignment on the probability of each preference category. Robust (HC0) standard errors in parentheses. $N = 390$. $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$.

Table A.5: 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 D.

Table A.6: Gelbach decomposition of the effect on demand for human therapy, adding behavioral activation as a mediator

Component	Contribution	SE	% of treatment effect
Explained by improved mental health symptoms	−0.035		28.1%
		(0.018)	
Explained by increased behavioral activation	−0.000		0.2%
		(0.012)	
via Preferences (all 3 dummies)	−0.027		21.4%
		(0.010)	
of which: AI therapy preference	−0.032		25.7%
		(0.011)	
of which: no preference	+0.006		-5.1%
		(0.005)	
of which: prefers no therapy	−0.001		0.8%
		(0.005)	
Residual direct effect	−0.062		50.3%
		(0.032)	
Total treatment effect (= sum of above)	−0.124		100.0%
		(0.030)	

Notes: Gelbach (2016) decomposition of the intention-to-treat effect of assignment to the AI app on stated demand for human therapy (Wave 5, 0–1 scale), adding behavioral activation as a third mediator alongside the mental-health composite and the therapy-preference dummies. All models, including the no-mediator baseline in the bottom row, are estimated on the $N = 386$ participants with non-missing mediators. Controls: baseline PHQ-8, GAD-7, prior beliefs about the efficacy of human therapy, prior therapy experience, and medication history. The mental-health channel is a standardized composite of Wave-5 GAD-7 and PHQ-8; behavioral activation is the standardized Wave-5 behavioral-activation score; the preference channel comprises dummies for preferring an AI therapist, having no preference, and preferring no therapy, each relative to preferring a human therapist (omitted). Each contribution is $\hat{\pi}_k \hat{\gamma}_k$, where $\hat{\pi}_k$ is the OLS treatment effect on mediator k and $\hat{\gamma}_k$ its partial effect on demand in the full model; standard errors (in parentheses) use the Gelbach (2016) GMM delta-method estimator. Behavioral activation contributes essentially nothing over and above the mental-health channel.

Table A.7: Engagement with the AI app over the four-week trial (treated arm)

	Mean	Median	p25	p75	Max
Number of sessions	15.4	13.0	5.0	22.0	77
Active days	10.5	9.0	3.0	17.0	29
Conversation hours	3.1	1.8	0.5	4.4	22

Notes: Per-participant engagement among participants assigned to the AI app, over the 28 days following randomization. Table A.8 reports the same measures cumulated over shorter windows. "Number of sessions" and "Active days" are summed from the application's daily logs; "Conversation hours" is cumulative voice-conversation time. Columns report the mean, median, 25th and 75th percentiles, and maximum across treated participants.

Table A.8: Engagement with the AI app by horizon (treated arm)

	Mean	SD	Median	p25	p75
First week					
Number of sessions	5.3	5.0	5.0	2.0	7.0
Active days	3.5	2.4	3.0	1.0	6.0
Conversation hours	1.09	1.22	0.69	0.20	1.53
First two weeks					
Number of sessions	9.4	8.0	8.0	4.0	14.0
Active days	6.3	4.4	5.5	2.0	10.0
Conversation hours	1.91	2.09	1.17	0.38	2.96
Full trial					
Number of sessions	15.4	13.3	13.0	5.0	22.0
Active days	10.5	8.1	9.0	3.0	17.0
Conversation hours	3.14	3.75	1.79	0.50	4.37

Notes: Per-participant engagement among the 200 participants assigned to the AI app, cumulated from randomization over three windows: the first week (days 1–7), the first two weeks (days 1–14), and the full trial (days 1–28). Measures are constructed from the application's server logs, as described in Section V.A of the main text. Sessions and active days are summed over the window; conversation hours are the total duration of spoken exchange with the application. Engagement is a pre-registered primary outcome. Compensation beyond the first session was tied to survey completion rather than to use of the application, so these figures reflect voluntary use. The full-trial rows repeat Table A.7 in a different set of statistics.

Table A.9: Main results, Wave 4

	GAD-7 (Points)	GAD-7 (Std.)	PHQ-8 (Points)	PHQ-8 (Std.))
Treatment	-3.52***	-0.63***	-3.03***	-0.67***
	(0.46)	(0.08)	(0.49)	(0.11)
Baseline Controls	✓	✓	✓	✓
Obs	390	390	390	390
Adj. R^2	0.22	0.22	0.35	0.35

Notes: This table reports OLS estimates of the effect of assignment to the AI app on mental health outcomes in Wave 4. Dependent variables are the GAD-7 and PHQ-8 scores, measured in points and in standardized form. Standardization is based on the full baseline sample. Baseline controls are the baseline mental health measures (Wave 1 GAD-7 or Wave 2 PHQ-8, respectively), an indicator for prior therapy history, and the participant's estimate of how effective human therapy is. Standard errors, reported in parentheses, are robust to heteroskedasticity. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table A.10: Qualitative coding scheme for barriers to human therapy

Human label	**Example quotes**
Therapist mismatch	Because I can't find a good therapist; I couldn't find a connection with a therapist; difficulty finding a therapist that I can relate to.
Access barriers	Just haven't gotten around to finding one that takes my insurance; still searching for one; unable to find one.
Perceived ineffectiveness	Because I have been to so many therapists in the past, and went to them for so many years, and in the end every one of them told me that they couldn't help me; I tried a human therapist not long after my mom passed away and she only let me talk for about 10 minutes of a 45-minute session.
Cost	Can't afford it; it's not something I could afford or something my parents would want to invest in.
Discomfort with human interaction	Hate telling my problems to people; hard opening up to anyone and not knowing how to talk about feelings or trauma; it scares me, I have social anxiety.
Time constraints	Lack of time and energy; takes up too much time.
Low energy	Lack of time and energy.
No perceived need	Self-reliance / not needed yet.
Stigma barriers	Negative stigma in family.
Prefer digital help	I feel AI understands me better.

Notes: Table reports the qualitative coding scheme used to classify open-text responses explaining why participants were not currently in human therapy. Labels were developed inductively and refined iteratively using example quotes from the survey.

Table A.11: Heterogeneity in the estimated effect of assignment

	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. The variable of interest is an indicator for assignment to the AI app. Heterogeneity is tested via interactions between assignment 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. *$p < 0.10$, **$p < 0.05$, ***$p \leq 0.01$

Table A.12: Heterogeneity in the estimated effect of assignment (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 Negativity 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

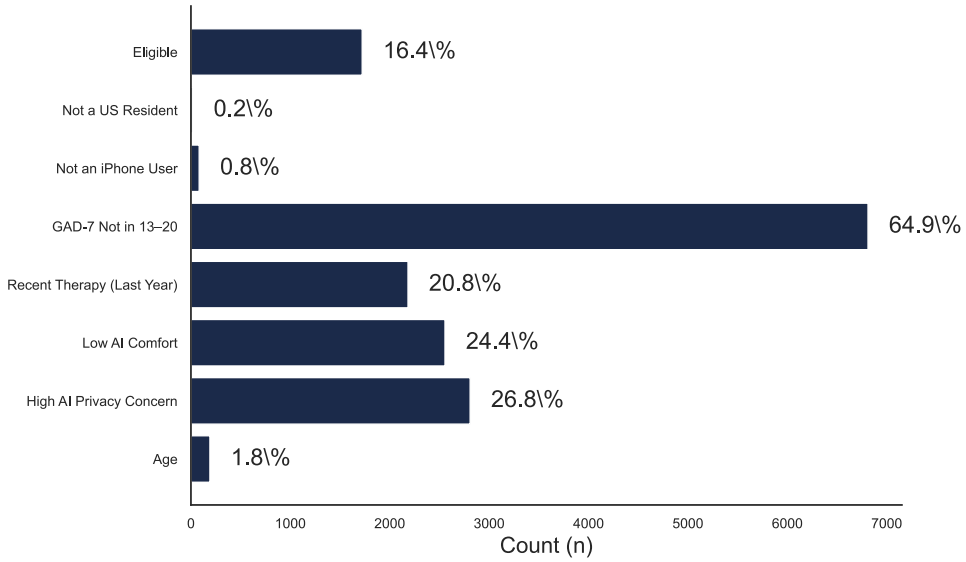
Notes: The dependent variable is standardized GAD-7 score at Wave 4. Each column reports OLS estimates with HC3 heteroskedasticity-robust standard errors in parentheses. The variable of interest is an indicator for assignment to the AI app. Heterogeneity is tested via interactions between assignment and baseline characteristics: above-median mental health severity (W1), GAD-7 (W1), PHQ-8 (W2), ATQ (W2), behavioral activation (W2), negativity bias (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. $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

B Additional Figures



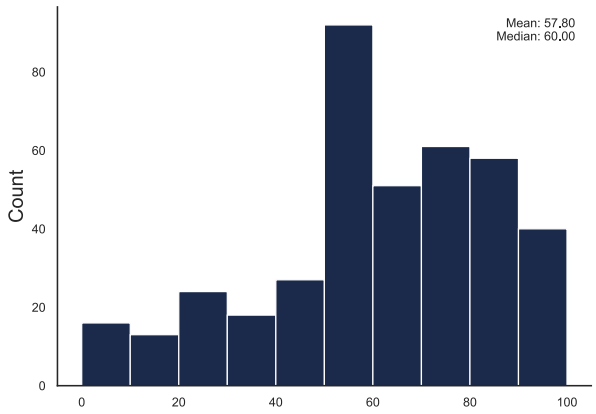
Figure B.1: Recruitment advertisement on Facebook and Instagram

Figure B.2: Screening failures among Wave 1 respondents



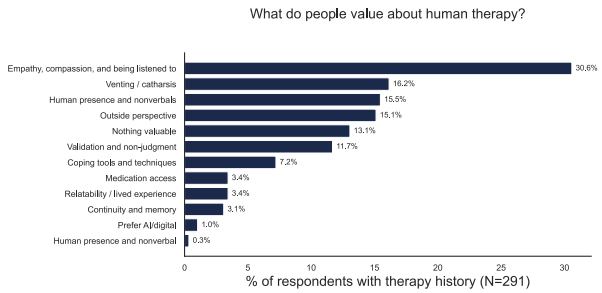
Note: This figure reports screening failures among all 10,490 participants completing the Wave 1 baseline questionnaire, administered immediately following advertisement click-through. Because failure categories are nonexclusive, cumulative failure percentages exceed 100%.

Figure B.3: Baseline beliefs about the efficacy of in-person therapy with a licensed psychologist (Wave 2)

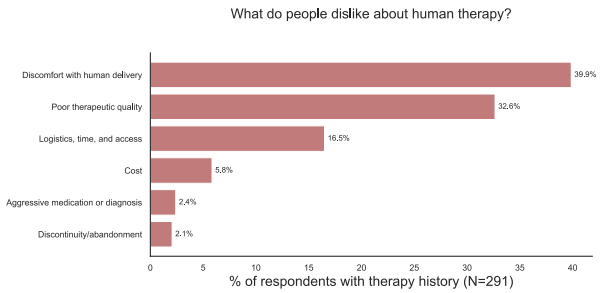


Note: Distribution of responses to: "If you did in-person therapy with a licensed psychologist, what do you think is the probability that your mental health would improve?", asked at baseline (Wave 2).

Figure B.4: Valued and disliked aspects of prior human therapy among respondents with therapy experience ($N = 291$)

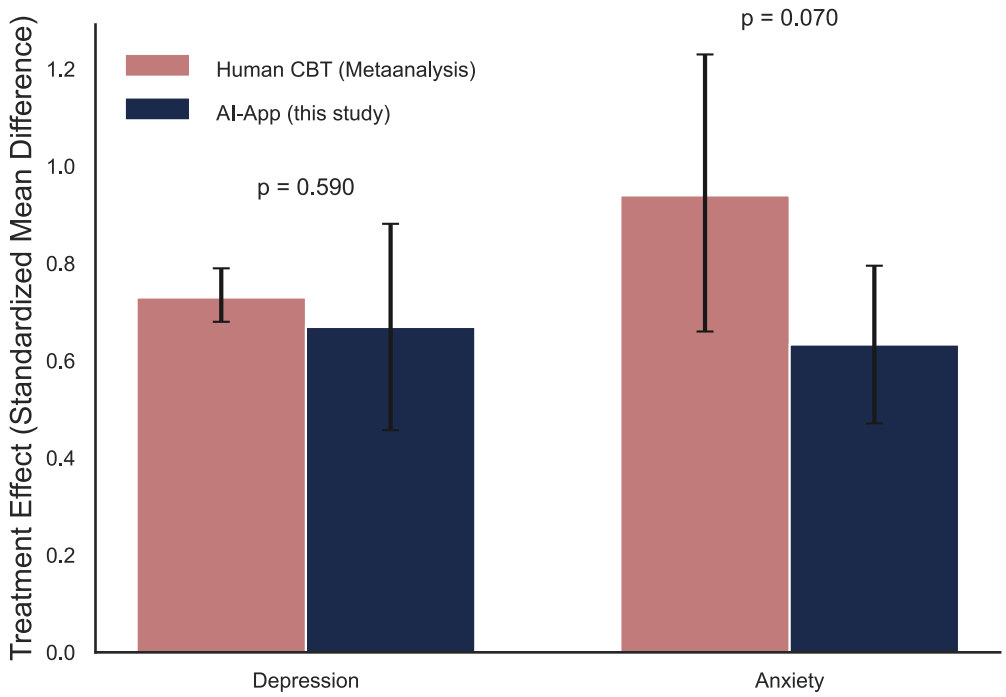


(a) What respondents value about human therapy.



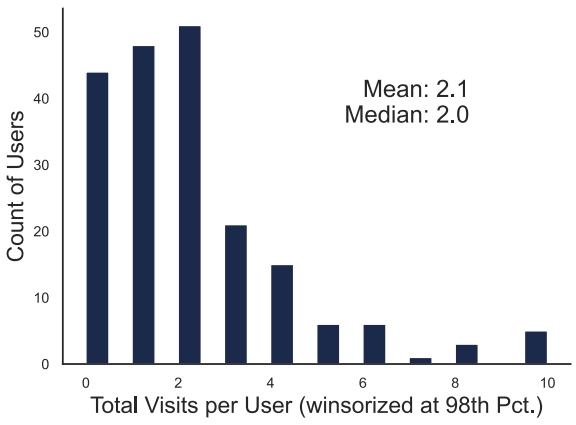
(b) What respondents dislike about human therapy.

Figure B.5: Standardized estimates in the present study versus meta-analytic benchmarks from Harrer et al. (2022)

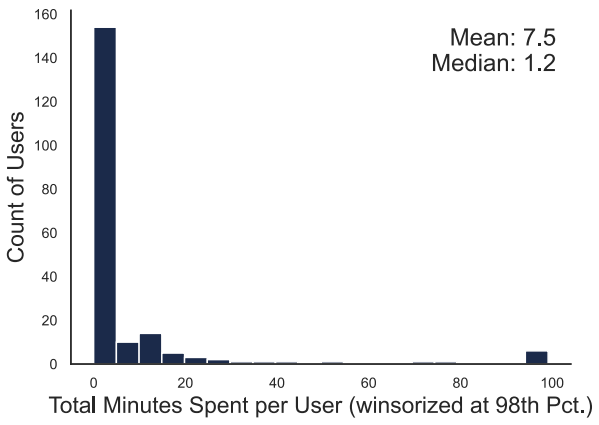


Note: Points show standardized mean difference estimates with 95% confidence intervals (CIs). The 4-week standardized effect on anxiety symptoms (GAD-7) observed in the present study is plotted alongside a MetaPsy benchmark for individual-format CBT versus waitlist/placebo at 6–8 weeks (dataset v23.0.4; 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).

Figure B.6: Engagement with the passive-control psychoeducation website



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

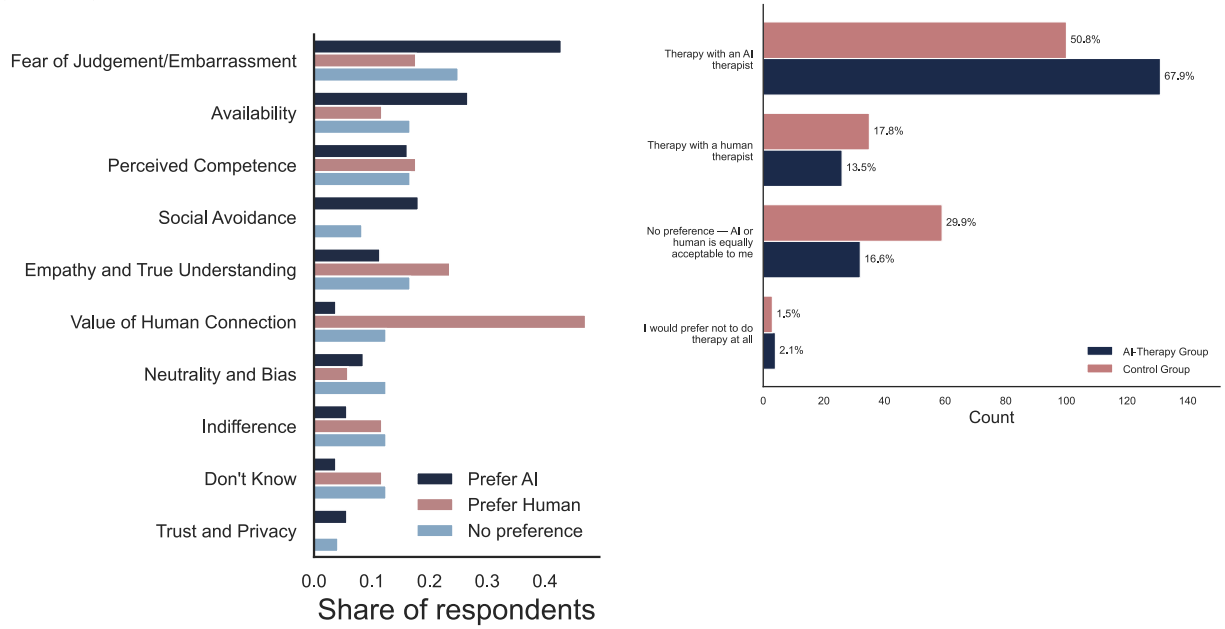


(b) Total minutes spent per user on web app (winsorized at the 98th percentile).

Figure B.7: Preferences for AI versus human therapy

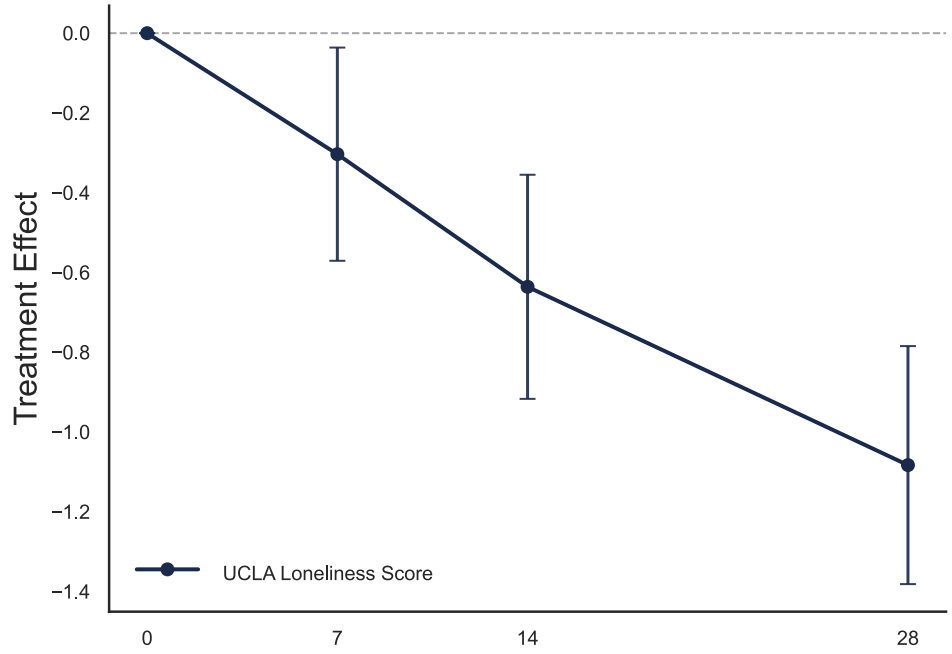
(a) Themes in AI vs. human therapy preferences (Wave 3).

(b) Stated therapy modality preference by arm (Wave 4).



Note: (a) Coded themes among respondents with prior therapy experience (Wave 3). Thematic categories were coded from Wave 3 responses using ten predefined categories; bars show the distribution of coded themes describing preferences between AI-based and human therapy. We display only categories occurring at least 15 times. (b) Stated therapy modality preference by arm at the two-week follow-up (Wave 4): distribution of respondents' preference for receiving therapy from an AI therapist versus a human therapist.

Figure B.8: Estimated effects of assignment to the AI app on loneliness (UCLA Loneliness scale)



Note: Points plot estimated intention-to-treat effects of assignment to the AI app on UCLA Loneliness at each follow-up wave, where the dependent variable is the standardized (z-scored) 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 B.9: Histograms of session lengths and engaged days in the AI-app arm

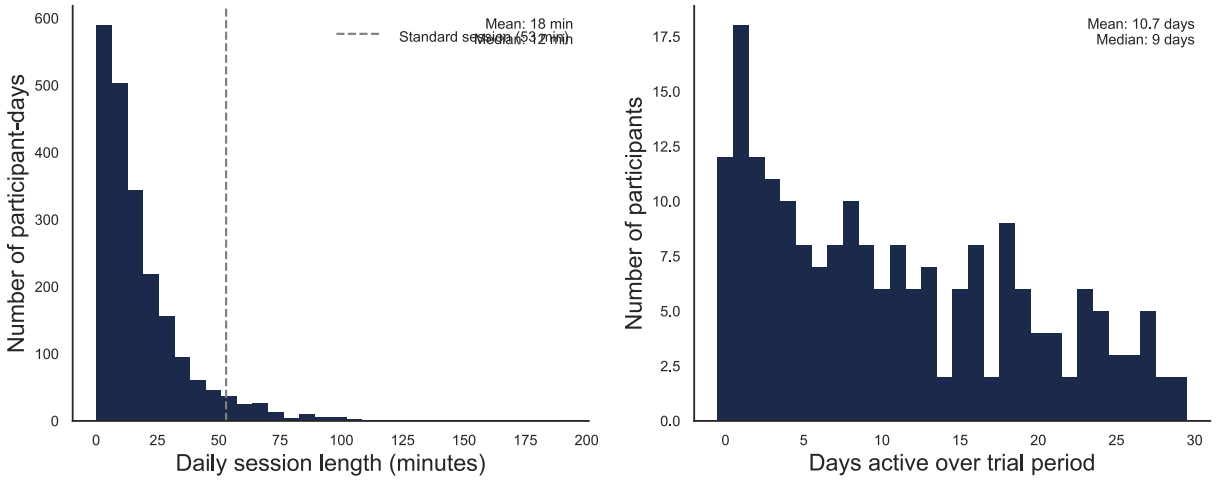
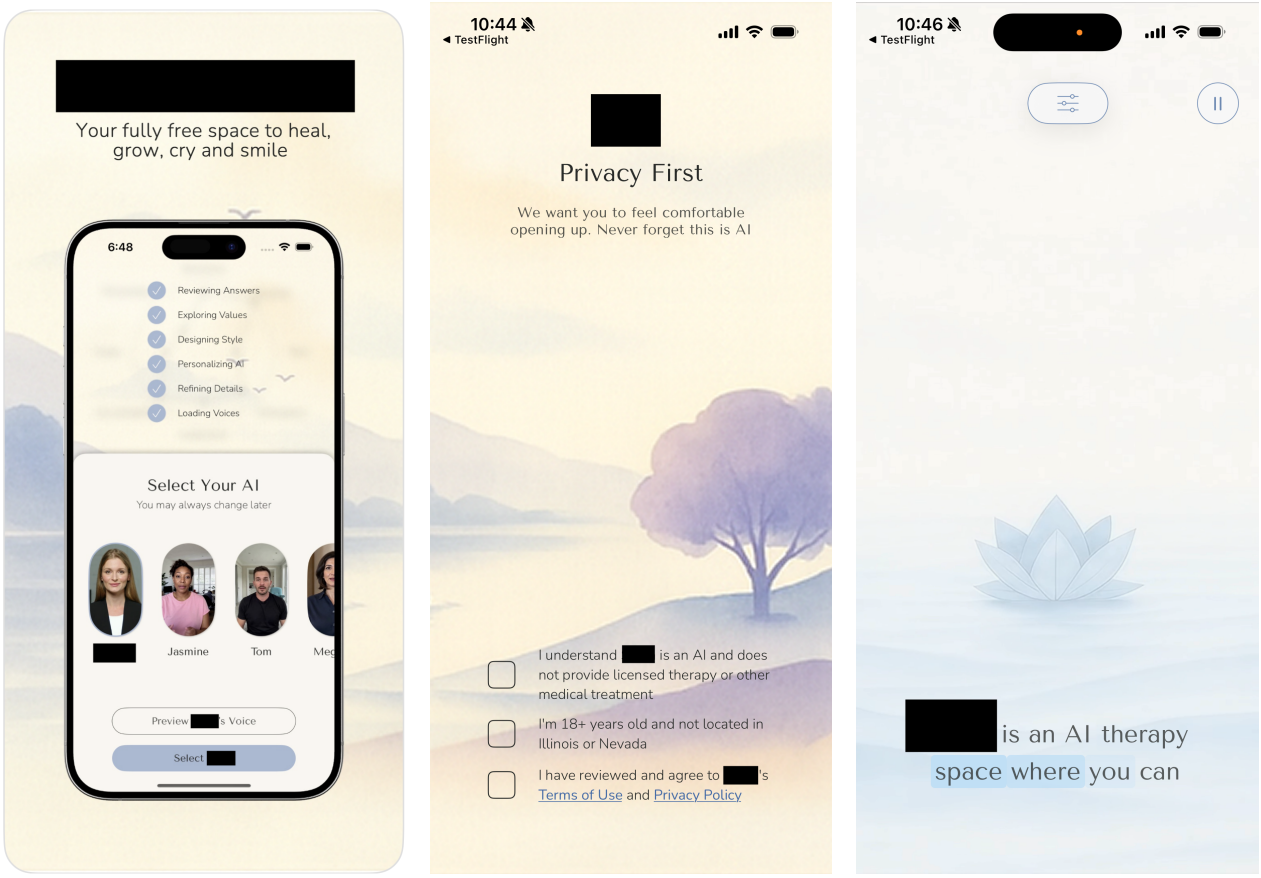


Figure B.10: Representative screenshots from the voice-based AI app



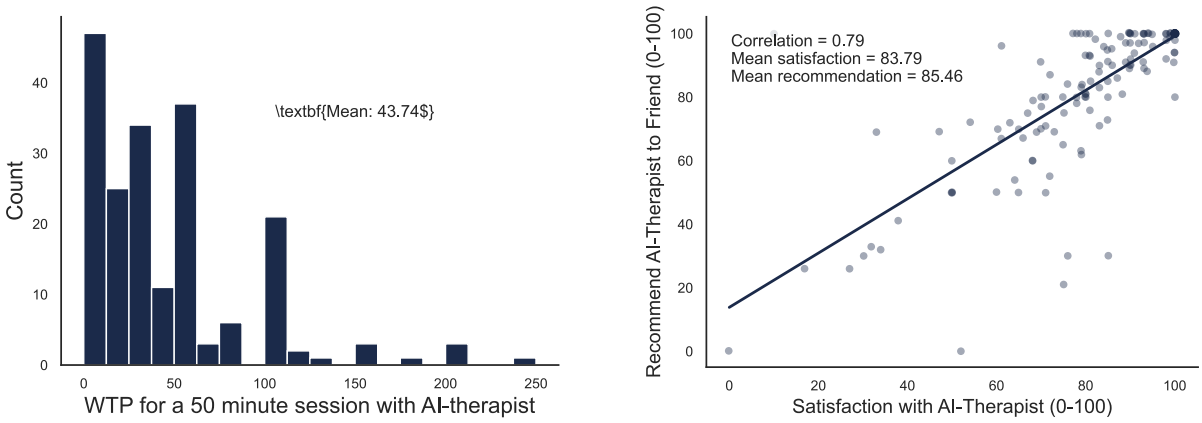
(a) Personalization options (e.g., voice and interaction style) used to tailor the AI therapist.

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

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

Note: The panels illustrate three key user-facing elements: personalization settings (left), onboarding privacy disclosure (middle), and the therapy-session interface (right). Reductions cover the tool's commercial brand name.

Figure B.12: User valuation and endorsement of the AI therapist

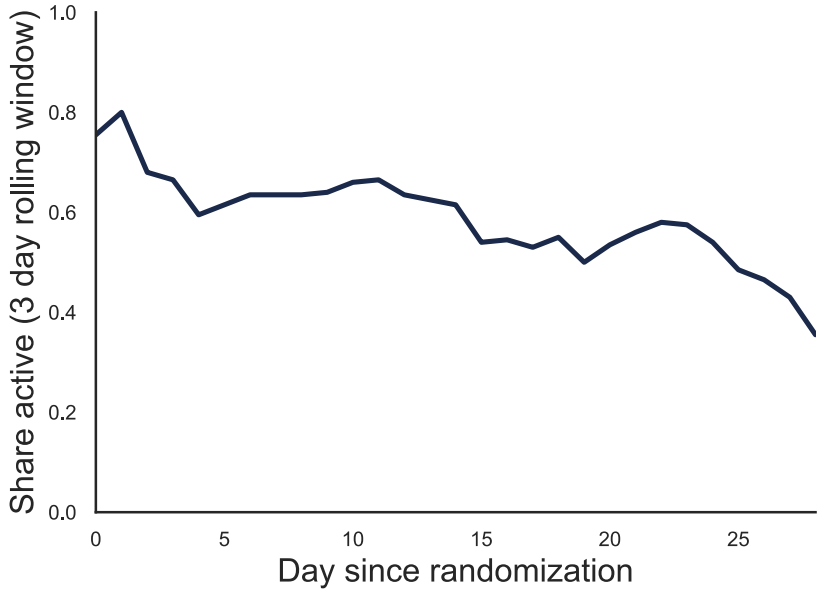


(a) Willingness to pay for a 50-minute session with the AI therapist (participants assigned to the AI app).

(b) Satisfaction and recommendation (Wave 3).

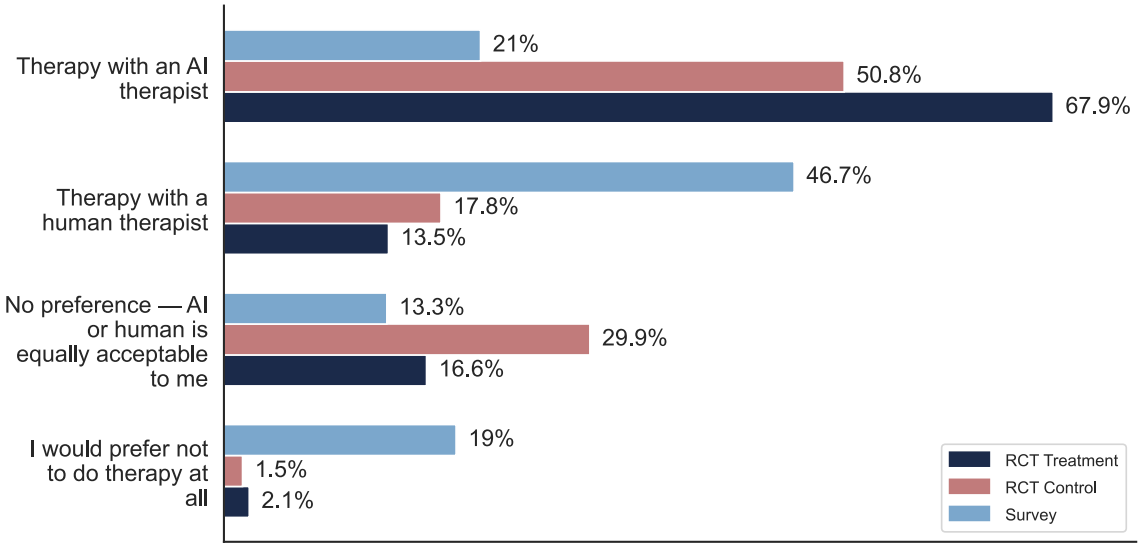
Note: Left panel: distribution of stated willingness to pay for a 50-minute AI-therapist session, with the sample mean annotated. Right panel: jittered scatter of satisfaction levels (0–100) with the AI therapist versus probability of recommendation to a friend or family member (0–100), with fitted regression line; panel reports correlation and sample means.

Figure B.11: Engagement over time in the AI-app arm



Note: 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 B.13: Overall preference for AI versus human therapy in the RCT and survey samples



Note: 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.

C 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 one 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:

Hi, I'm [AI voice name], 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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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:

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.

D 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 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).

E Statistical Analysis and Deviations from the Pre-Registration

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

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

where Y_i denotes the outcome for participant i at endline, Treatment_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 are heteroskedasticity-robust and clustered at the individual level. For Wave 2 outcomes without baseline values, we control for baseline GAD-7. Standardization is done using the full baseline survey sample for GAD-7 ($N = 10,490$) and using the full randomized sample for PHQ-8 ($N = 400$).

Due to a technical error, we did not collect the pre-registered baseline covariate for motivation to work on one's mental health; this covariate was included to improve statistical power rather than to control for a confounding variable, and we show robustness to including several proxies for motivation in Section F. Survey responses were included if submitted within 5 days of the scheduled due date (we did not automatically close the survey; only one observation required this extension).

We document all deviations from the pre-registration in the table below.

Table E.1: Deviations from the pre-registration

Pre-registered specification	**Deviation**	**Reason**
Include hmotivation to work on one's mental health (Wave 2, 0–100 scale) as a baseline covariate in X_i.	Dropped this covariate from the analysis.	A technical error meant the item was not collected at baseline. It was included to improve statistical power rather than to address confounding; we show robustness to including various proxies for motivation in Section F.
Standard survey-wave timing; no rule pre-specified for late submissions.	Included survey responses submitted within 5 days of the scheduled due date.	The survey was not automatically closed at the due date; in practice only one observation required this extension.
Exclude participants who fail minimum engagement criteria in the baseline survey ("extremely short completion times"); no threshold specified.	Operationalized as excluding participants who spent fewer than 20 seconds completing the GAD-7.	The pre-registration did not specify a numeric threshold; 20 seconds was chosen to exclude bots and inattentive responses while retaining attentive participants.
Social outcomes were limited to loneliness (3-item UCLA Loneliness Scale); the social-contact items were not pre-specified.	Added items on the frequency of meaningful social interaction and of seeking emotional support from friends and family (Wave 5).	These items were introduced during the trial, after we observed how much time participants spent in conversation with the AI app, to test whether this engagement crowded out human social contact (our "social crowding-in" analysis).
Elicit interest in human-led therapy on a 5-point Likert scale.	Measured demand for human therapy on a 0–100 scale instead.	A continuous 0–100 scale gives finer resolution and matches the format of the medication-inclination and treatment-belief items, so the demand measures can be analyzed and standardized on a common scale.
Elicit beliefs in the effectiveness of psychiatric medication alongside human-led and AI-based support.	The medication-effectiveness belief item was not fielded in any wave.	Due to an oversight. The registered human-therapy and AI effectiveness beliefs are reported in Table A.2.

Notes: This table summarizes deviations from the pre-registered analysis plan (AsPredicted #258,203). Rows 1–2 are also discussed in the text of Section E.

Table E.2: Pre-registered outcomes and analyses (AsPredicted #258203) and where they are reported

Pre-registered outcome / analysis	Type	Reported in
Panel A: Primary outcomes		
Standardized GAD-7 change at two weeks (Wave 4)	Primary	Companion paper (Röver et al., 2026); all-wave extension in Section VI of the main text
Self-reported safety (open-ended, Wave 4)	Primary	Companion paper (Röver et al., 2026)
Engagement (sessions, exchanges, active days, usage duration)	Primary	Section V.A of the main text; results in Figure Va
User satisfaction (satisfaction, WTP, willingness to recommend)	Primary	Section VI.B of the main text (Figure B.12)
Panel B: Secondary and exploratory analyses		
Standardized PHQ-8 change	Secondary	Companion paper (two-week) (Röver et al., 2026); all waves in Section VI of the main text
Composite mental-health index (standardized GAD-7 + PHQ-8)	Secondary	Section VI of the main text
Complier average causal effect (CACE) on GAD-7	Secondary	Appendix F.2 (Table F.14)
Demand for human-led and AI-based support	Secondary	Section VII.A of the main text (Table II)
WTP for human-led support (effect); AI WTP (treated mean)	Secondary	Table A.3; Figure B.12a
Beliefs in effectiveness of human-led and AI-based support (medication-effectiveness belief not fielded; Table E.1)	Secondary	Table A.2
Medication inclination (0–100); interest in human therapy (Likert); teletherapy-link click	Secondary	Section VII.A of the main text (Table II)
CBT mechanisms (automatic thoughts, behavioral activation, negativity bias), all waves	Secondary	Section VI.B of the main text
Loneliness (3-item UCLA Scale)	Secondary	Section VI.B of the main text
Effect of AI-information treatment on take-up (Wave 1 cross-randomization)	Secondary	Appendix J
Heterogeneity by mechanisms, cross-randomization, and engagement	Secondary (underpowered)	Appendix G
Experimenter-demand and social-desirability checks	Secondary	Appendix F

Notes: The table maps every outcome and analysis pre-registered under AsPredicted #258203 to its location across the two-paper research program. Primary outcomes on symptoms and safety are reported in the companion paper (Röver et al., 2026); this paper reports the pre-registered secondary outcomes and pre-specified extensions. Analyses not listed here, or listed as heterogeneity analyses, are exploratory and labelled as such in the text. The composite mental-health index averages the separately standardized GAD-7 and PHQ-8 scores, following the pre-registration.

F Additional Robustness

In this section we subject every treatment effect reported in the main body to the same battery of robustness checks. For each effect we report one consolidated table whose rows are the robustness checks and whose columns are the follow-up wave(s): the main specification; the same specification adding one baseline covariate at a time (willingness to pay for human therapy; baseline beliefs about the AI's efficacy, emotional comfort, and privacy; and household income); controls for self- and social-image concerns and a subsample dropping respondents who report feeling pressured to adjust their answers (social desirability and experimenter demand); a conservative baseline-carried-forward imputation for attriters; and Lee (2009) trimming bounds (Lee, 2009). Symptom, mechanism, loneliness, and social-interaction effects are reported at the two-week follow-up (Wave 4); the demand effects, which we track over time, are reported at every wave at which they are measured. We first discuss the estimated symptom effects (Section F.1), and then the complier average causal effect (Section F.2).

F.1 Robustness for the Estimated Treatment Effects

This section evaluates whether the estimated benefits of assignment to the AI app 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 The estimated effects are virtually unchanged across alternative specifications that add different baseline covariates, including income and medication use, which are slightly imbalanced across groups at baseline. For the symptom outcomes this is shown in the covariate rows of Tables F.2 and F.3; the same pattern holds for every other outcome in the corresponding per-effect tables below. Point estimates and standard errors move only minimally, suggesting that our estimates are not sensitive to specification choice or residual baseline imbalances.

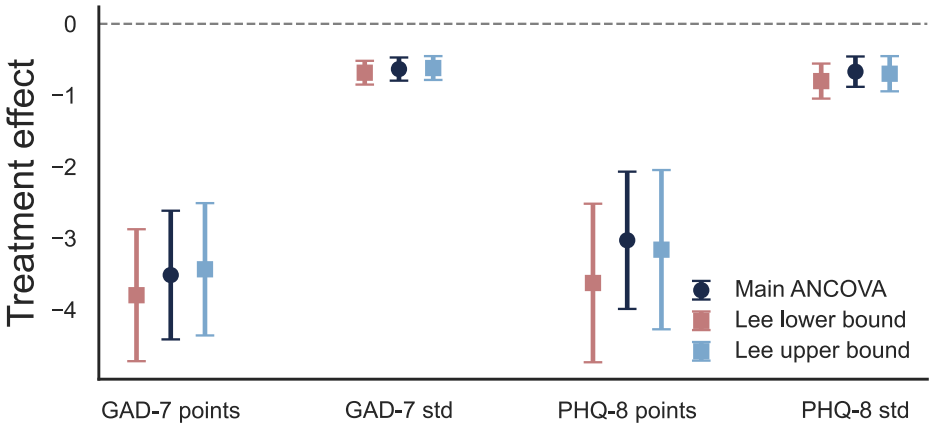
Attrition Attrition by the 2-week follow-up (where we measure our primary GAD-7 outcome) is only 2.5% (10 participants), which substantially limits scope for attrition bias (Table F.1). Under a conservative baseline-carried-forward imputation for attriters, our estimates remain essentially identical to the main specification (the "Attrition: baseline carry-forward" row of each per-effect table), and Lee bounds are tightly clustered around the point estimates (the "Lee bounds" row; see also Figure F.1), indicating that selective attrition is unlikely to account for the findings (Lee, 2009).

Table F.1: 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 AI-app and control arms across survey waves. Wave 1 was conducted prior to randomization and therefore represents the pre-intervention sample. Subsequent waves are post-randomization follow-ups. Counts reflect completed surveys in each wave.

Figure F.1: Main ANCOVA estimates and Lee bounds for the estimated effect of assignment on anxiety (GAD-7) and depression (PHQ-8)



Note: The figure plots intention-to-treat estimates of the effect of assignment to the AI therapy intervention versus control, obtained from the pre-registered ANCOVA specification (endline outcome regressed on 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 estimated effect.

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 our estimates only if the intervention differentially affects individuals' willingness to acknowledge poor mental health, that is, if self-image concerns differ across arms after the intervention. We follow Bursztyn et al. (2025) and assess self-image concerns around mental health in Wave 5. Figure F.2a shows the answer distributions across arms; the p-value of a Mann–Whitney U-test is 0.16, alleviating concerns that differential SDB across arms contaminates our estimates. We also re-estimate our main specification controlling for self-image concerns, and results remain robust (the "+ Self-image control" row of each per-effect table).

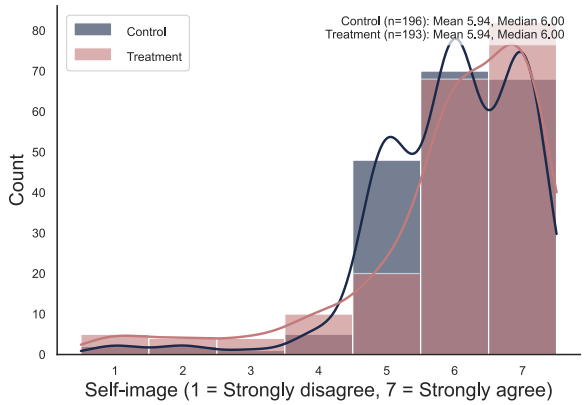
Experimenter Demand Another potential concern is experimenter demand effects, which in this context constitute a specific form of SDB: participants in the AI-app arm may infer that the experimental hypothesis predicts improvement and adjust their responses to align with the perceived expectations of 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 Figures B.6a and B.6b). 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.

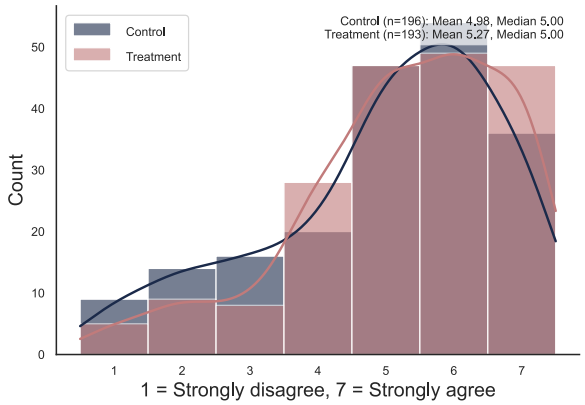
We explicitly test for experimenter demand effects in Wave 5, following best practices from the literature (Bursztyn et al., 2025; De Quidt et al., 2018). Participants are asked whether they felt pressured to adjust their responses during the trial and about social-image concerns related to mental health (Figure F.2b). We then re-estimate our main specification controlling for social-image concerns (the "+ Social-image control" row of each per-effect table) and excluding participants who report feeling pressured to adjust their answers (the "Drop experimenter-demand resp." row). In both cases, the results remain virtually unchanged, alleviating concerns that experimenter demand effects drive our findings.

Figure F.2: Self-image and social-image response distributions across arms (Wave 5)

(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 $p = 0.13$.



The Full Battery for Every Effect We apply exactly the same checks to the remaining main-body outcomes: the CBT mechanisms (automatic thoughts, behavioral activation, and negativity bias), perceived loneliness, the social-interaction items, and the demand outcomes.[1] Across all checks, the estimated treatment effects retain their sign and significance. The consolidated per-effect tables follow.

Robustness of the Social-Contact Heterogeneity Section VII.C of the main text reports that the treatment effect on social contact concentrates among participants at the maximum of the baseline loneliness scale. Table F.11 probes the interaction coefficient directly. Because treatment is randomized, additively entered baseline covariates cannot move an interaction between treatment and a baseline trait; the checks that bind are competing interactions and differential reporting. The interaction retains its magnitude and significance when treatment is additionally interacted with baseline anxiety or depression, when the Wave-5 self- and social-image items enter both directly and interacted with treatment, and when respondents who report pressure to adjust their answers are dropped. Attrition is unrelated to the treatment-by-loneliness cell.

The same interaction applied to the loneliness outcome itself shows no heterogeneity: the effect

[1]The baseline-carried-forward check requires a pre-treatment measure of the same item and is therefore reported only for the mechanism and loneliness outcomes; the demand, online-interest, and social-interaction items were first elicited after randomization, so no pre-treatment baseline exists for them ("—" in those rows). Lee bounds are likewise not defined for the binary-logit columns.

Table F.2: Robustness battery for the estimated effect on GAD-7 (Std.).

	Week 1	Week 2	Week 4
Main specification	-0.33*** (0.07), $N = 393$	-0.63*** (0.08), $N = 390$	-0.81*** (0.09), $N = 390$
+ WTP for human therapy	-0.34*** (0.07), $N = 385$	-0.66*** (0.08), $N = 382$	-0.82*** (0.09), $N = 382$
+ AI efficacy belief	-0.33*** (0.07), $N = 393$	-0.63*** (0.08), $N = 390$	-0.81*** (0.09), $N = 390$
+ AI emotional comfort	-0.33*** (0.07), $N = 393$	-0.64*** (0.08), $N = 390$	-0.81*** (0.09), $N = 390$
+ AI privacy	-0.33*** (0.07), $N = 393$	-0.63*** (0.08), $N = 390$	-0.81*** (0.09), $N = 390$
+ Household income	-0.36*** (0.07), $N = 385$	-0.67*** (0.08), $N = 382$	-0.84*** (0.09), $N = 382$
+ Self-image control	-0.32*** (0.08), $N = 387$	-0.64*** (0.08), $N = 386$	-0.83*** (0.09), $N = 389$
+ Social-image control	-0.29*** (0.07), $N = 387$	-0.61*** (0.08), $N = 386$	-0.78*** (0.09), $N = 389$
Drop experimenter-demand resp.	-0.31*** (0.07), $N = 374$	-0.61*** (0.08), $N = 373$	-0.77*** (0.09), $N = 376$
Attrition: baseline carry-forward	-0.31*** (0.07), $N = 400$	-0.59*** (0.08), $N = 400$	-0.77*** (0.09), $N = 400$
Lee bounds [lower, upper]	[-0.39, -0.33]	[-0.68, -0.62]	[-0.84, -0.80]

Notes: Each cell reports the estimated treatment effect on the column outcome, its standard error in parentheses, and the sample size N. OLS / linear-probability estimates use HC3 robust standard errors; ordered- and binary-logit estimates (in log-odds) use model-based standard errors. Rows report robustness for the treatment coefficient: the main specification; then adding one baseline covariate at a time (willingness to pay for human therapy; beliefs about AI efficacy, emotional comfort, and privacy; household income); controlling for self-image and social-image concerns (Wave-5 self-reports); dropping respondents who report feeling pressured to adjust their answers; baseline carry-forward for attriters; and Lee (2009) trimming bounds. Carry-forward and Lee bounds are shown only where applicable ("—" otherwise). $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table F.3: Robustness battery for the estimated effect on PHQ-8 (Std.).

	Week 1	Week 2	Week 4
Main specification	-0.30*** (0.10), $N = 393$	-0.67*** (0.11), $N = 390$	-0.96*** (0.11), $N = 390$
+ WTP for human therapy	-0.31*** (0.10), $N = 385$	-0.71*** (0.11), $N = 382$	-1.00*** (0.11), $N = 382$
+ AI efficacy belief	-0.30*** (0.10), $N = 393$	-0.67*** (0.11), $N = 390$	-0.96*** (0.11), $N = 390$
+ AI emotional comfort	-0.30*** (0.10), $N = 393$	-0.67*** (0.11), $N = 390$	-0.96*** (0.11), $N = 390$
+ AI privacy	-0.30*** (0.10), $N = 393$	-0.67*** (0.11), $N = 390$	-0.96*** (0.11), $N = 390$
+ Household income	-0.34*** (0.10), $N = 385$	-0.73*** (0.11), $N = 382$	-1.03*** (0.11), $N = 382$
+ Self-image control	-0.29*** (0.10), $N = 387$	-0.69*** (0.11), $N = 386$	-0.98*** (0.11), $N = 389$
+ Social-image control	-0.26** (0.10), $N = 387$	-0.65*** (0.11), $N = 386$	-0.94*** (0.11), $N = 389$
Drop experimenter-demand resp.	-0.28*** (0.10), $N = 374$	-0.69*** (0.11), $N = 373$	-0.93*** (0.11), $N = 376$
Attrition: baseline carry-forward	-0.28*** (0.10), $N = 400$	-0.63*** (0.11), $N = 400$	-0.92*** (0.11), $N = 400$
Lee bounds [lower, upper]	[-0.42, -0.34]	[-0.80, -0.70]	[-1.04, -0.99]

Notes: Each cell reports the estimated treatment effect on the column outcome, its standard error in parentheses, and the sample size N. OLS / linear-probability estimates use HC3 robust standard errors; ordered- and binary-logit estimates (in log-odds) use model-based standard errors. Rows report robustness for the treatment coefficient: the main specification; then adding one baseline covariate at a time (willingness to pay for human therapy; beliefs about AI efficacy, emotional comfort, and privacy; household income); controlling for self-image and social-image concerns (Wave-5 self-reports); dropping respondents who report feeling pressured to adjust their answers; baseline carry-forward for attriters; and Lee (2009) trimming bounds. Carry-forward and Lee bounds are shown only where applicable ("—" otherwise). $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table F.4: Robustness battery for the estimated effect on Automatic Thoughts.

	Week 1	Week 2	Week 4
Main specification	-0.30*** (0.08), $N = 393$	-0.55*** (0.09), $N = 390$	-0.84*** (0.09), $N = 390$
+ WTP for human therapy	-0.32*** (0.08), $N = 385$	-0.58*** (0.09), $N = 382$	-0.86*** (0.10), $N = 382$
+ AI efficacy belief	-0.30*** (0.08), $N = 393$	-0.55*** (0.09), $N = 390$	-0.84*** (0.09), $N = 390$
+ AI emotional comfort	-0.30*** (0.08), $N = 393$	-0.55*** (0.09), $N = 390$	-0.84*** (0.09), $N = 390$
+ AI privacy	-0.30*** (0.08), $N = 393$	-0.55*** (0.09), $N = 390$	-0.84*** (0.09), $N = 390$
+ Household income	-0.32*** (0.08), $N = 385$	-0.58*** (0.09), $N = 382$	-0.88*** (0.10), $N = 382$
+ Self-image control	-0.30*** (0.08), $N = 387$	-0.54*** (0.09), $N = 386$	-0.86*** (0.09), $N = 389$
+ Social-image control	-0.27*** (0.08), $N = 387$	-0.51*** (0.09), $N = 386$	-0.80*** (0.09), $N = 389$
Drop experimenter-demand resp.	-0.30*** (0.08), $N = 374$	-0.53*** (0.09), $N = 373$	-0.82*** (0.09), $N = 376$
Attrition: baseline carry-forward	-0.29*** (0.08), $N = 400$	-0.53*** (0.09), $N = 400$	-0.81*** (0.09), $N = 400$
Lee bounds [lower, upper]	[-0.34, -0.28]	[-0.61, -0.52]	[-0.85, -0.81]

Notes: Each cell reports the estimated treatment effect on the column outcome, its standard error in parentheses, and the sample size N. OLS / linear-probability estimates use HC3 robust standard errors; ordered- and binary-logit estimates (in log-odds) use model-based standard errors. Rows report robustness for the treatment coefficient: the main specification; then adding one baseline covariate at a time (willingness to pay for human therapy; beliefs about AI efficacy, emotional comfort, and privacy; household income); controlling for self-image and social-image concerns (Wave-5 self-reports); dropping respondents who report feeling pressured to adjust their answers; baseline carry-forward for attriters; and Lee (2009) trimming bounds. Carry-forward and Lee bounds are shown only where applicable ("—" otherwise). $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table F.5: Robustness battery for the estimated effect on Behavioral Activation.

	Week 1	Week 2	Week 4
Main specification	0.26*** (0.10), $N = 393$	0.50*** (0.10), $N = 390$	0.79*** (0.11), $N = 390$
+ WTP for human therapy	0.27*** (0.10), $N = 385$	0.54*** (0.10), $N = 382$	0.82*** (0.11), $N = 382$
+ AI efficacy belief	0.26*** (0.10), $N = 393$	0.50*** (0.10), $N = 390$	0.79*** (0.11), $N = 390$
+ AI emotional comfort	0.26*** (0.10), $N = 393$	0.50*** (0.10), $N = 390$	0.79*** (0.11), $N = 390$
+ AI privacy	0.26*** (0.10), $N = 393$	0.50*** (0.10), $N = 390$	0.79*** (0.11), $N = 390$
+ Household income	0.30*** (0.10), $N = 385$	0.56*** (0.10), $N = 382$	0.85*** (0.11), $N = 382$
+ Self-image control	0.22** (0.10), $N = 387$	0.49*** (0.11), $N = 386$	0.77*** (0.11), $N = 389$
+ Social-image control	0.21** (0.09), $N = 387$	0.49*** (0.10), $N = 386$	0.75*** (0.11), $N = 389$
Drop experimenter-demand resp.	0.21** (0.09), $N = 374$	0.52*** (0.11), $N = 373$	0.77*** (0.11), $N = 376$
Attrition: baseline carry-forward	0.24*** (0.09), $N = 400$	0.47*** (0.10), $N = 400$	0.75*** (0.11), $N = 400$
Lee bounds [lower, upper]	[0.17, 0.23]	[0.41, 0.49]	[0.71, 0.76]

Notes: Each cell reports the estimated treatment effect on the column outcome, its standard error in parentheses, and the sample size N. OLS / linear-probability estimates use HC3 robust standard errors; ordered- and binary-logit estimates (in log-odds) use model-based standard errors. Rows report robustness for the treatment coefficient: the main specification; then adding one baseline covariate at a time (willingness to pay for human therapy; beliefs about AI efficacy, emotional comfort, and privacy; household income); controlling for self-image and social-image concerns (Wave-5 self-reports); dropping respondents who report feeling pressured to adjust their answers; baseline carry-forward for attriters; and Lee (2009) trimming bounds. Carry-forward and Lee bounds are shown only where applicable ("—" otherwise). $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table F.6: Robustness battery for the estimated effect on Social Contact Frequency.

	Week 4
Main specification	0.64^{***} (0.18), $N = 390$
+ WTP for human therapy	0.65^{***} (0.19), $N = 382$
+ AI efficacy belief	0.65^{***} (0.18), $N = 390$
+ AI emotional comfort	0.64^{***} (0.18), $N = 390$
+ AI privacy	0.64^{***} (0.18), $N = 390$
+ Household income	0.70^{***} (0.19), $N = 382$
+ Self-image control	0.65^{***} (0.19), $N = 390$
+ Social-image control	0.60^{***} (0.19), $N = 390$
Drop experimenter-demand resp.	0.67^{***} (0.19), $N = 376$
Attrition: baseline carry-forward	—
Lee bounds [lower, upper]	[0.57, 0.63]

Notes: Each cell reports the estimated treatment effect on the column outcome, its standard error in parentheses, and the sample size *N*. OLS / linear-probability estimates use HC3 robust standard errors; ordered- and binary-logit estimates (in log-odds) use model-based standard errors. Rows report robustness for the treatment coefficient: the main specification; then adding one baseline covariate at a time (willingness to pay for human therapy; beliefs about AI efficacy, emotional comfort, and privacy; household income); controlling for self-image and social-image concerns (Wave-5 self-reports); dropping respondents who report feeling pressured to adjust their answers; baseline carry-forward for attriters; and Lee (2009) trimming bounds. Carry-forward and Lee bounds are shown only where applicable ("—" otherwise). $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table F.7: Robustness battery for the estimated effect on Negativity Bias.

	Week 4
Main specification	-0.78*** (0.09), $N = 390$
+ WTP for human therapy	-0.80*** (0.09), $N = 382$
+ AI efficacy belief	-0.78*** (0.09), $N = 390$
+ AI emotional comfort	-0.78*** (0.09), $N = 390$
+ AI privacy	-0.78*** (0.09), $N = 390$
+ Household income	-0.81*** (0.09), $N = 382$
+ Self-image control	-0.77*** (0.09), $N = 389$
+ Social-image control	-0.76*** (0.09), $N = 389$
Drop experimenter-demand resp.	-0.78*** (0.09), $N = 376$
Attrition: baseline carry-forward	-0.77*** (0.09), $N = 400$
Lee bounds [lower, upper]	[-0.72, -0.69]

Notes: Each cell reports the estimated treatment effect on the column outcome, its standard error in parentheses, and the sample size *N*. OLS / linear-probability estimates use HC3 robust standard errors; ordered- and binary-logit estimates (in log-odds) use model-based standard errors. Rows report robustness for the treatment coefficient: the main specification; then adding one baseline covariate at a time (willingness to pay for human therapy; beliefs about AI efficacy, emotional comfort, and privacy; household income); controlling for self-image and social-image concerns (Wave-5 self-reports); dropping respondents who report feeling pressured to adjust their answers; baseline carry-forward for attriters; and Lee (2009) trimming bounds. Carry-forward and Lee bounds are shown only where applicable ("—" otherwise). $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table F.8: Robustness battery for the estimated effect on UCLA Loneliness.

	Week 4
Main specification	-0.70*** (0.10), $N = 389$
+ WTP for human therapy	-0.75*** (0.10), $N = 381$
+ AI efficacy belief	-0.70*** (0.10), $N = 389$
+ AI emotional comfort	-0.71*** (0.10), $N = 389$
+ AI privacy	-0.70*** (0.10), $N = 389$
+ Household income	-0.74*** (0.10), $N = 381$
+ Self-image control	-0.69*** (0.10), $N = 389$
+ Social-image control	-0.68*** (0.10), $N = 389$
Drop experimenter-demand resp.	-0.68*** (0.10), $N = 376$
Attrition: baseline carry-forward	-0.68*** (0.10), $N = 400$
Lee bounds [lower, upper]	[-0.71, -0.65]

Notes: Each cell reports the estimated treatment effect on the column outcome, its standard error in parentheses, and the sample size N. OLS / linear-probability estimates use HC3 robust standard errors; ordered- and binary-logit estimates (in log-odds) use model-based standard errors. Rows report robustness for the treatment coefficient: the main specification; then adding one baseline covariate at a time (willingness to pay for human therapy; beliefs about AI efficacy, emotional comfort, and privacy; household income); controlling for self-image and social-image concerns (Wave-5 self-reports); dropping respondents who report feeling pressured to adjust their answers; baseline carry-forward for attriters; and Lee (2009) trimming bounds. Carry-forward and Lee bounds are shown only where applicable ("—" otherwise). $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

of assignment on UCLA loneliness does not differ detectably by baseline loneliness at any wave (Table F.12).

As a descriptive supplement, Table F.13 interacts treatment with the Wave-3 classification of whether a participant's open-ended explanation of their provider preference mentions fear of judgment or social avoidance. This split is not causal moderation: the classification is measured after randomization, less fear of judgment is part of what the treatment produces, and the underlying item differs across arms (treated participants explain their stated preference after app use; control participants answer a counterfactual comparison question), so the interaction carries a bad-control component. Neither interaction is significant.

F.2 Complier Average Causal Effect

Our main estimates are intention-to-treat effects of assignment to the AI app. Because take-up was not universal, we also report the complier average causal effect (CACE), the effect on those who actually engage with the intervention. We define compliance as completing at least one genuine AI session (a day with at least ten minutes of conversation); 83.5% of assigned participants comply, and non-compliance is one-sided, as control participants have no access to the app. The CACE is then the Wald ratio of the ITT to the compliance rate. Table F.14 reports both. Because compliance is

Table F.9: Robustness battery: demand for human therapy and online-therapy sign-up.

	Therapy Week 1	Therapy Week 4	Online sign-up (LPM, Week 4)	Online sign-up (Logit, Week 4)
Main specification	-0.13*** (0.03), $N = 347$	-0.13*** (0.03), $N = 389$	-0.18*** (0.05), $N = 390$	-0.76*** (0.21), $N = 390$
+ WTP for human therapy	-0.12*** (0.03), $N = 339$	-0.12*** (0.03), $N = 381$	-0.17*** (0.05), $N = 382$	-0.73*** (0.22), $N = 382$
+ AI efficacy belief	-0.13*** (0.03), $N = 347$	-0.13*** (0.03), $N = 389$	-0.18*** (0.05), $N = 390$	-0.77*** (0.21), $N = 390$
+ AI emotional comfort	-0.13*** (0.03), $N = 347$	-0.13*** (0.03), $N = 389$	-0.18*** (0.05), $N = 390$	-0.77*** (0.21), $N = 390$
+ AI privacy	-0.13*** (0.03), $N = 347$	-0.13*** (0.03), $N = 389$	-0.18*** (0.05), $N = 390$	-0.76*** (0.21), $N = 390$
+ Household income	-0.13*** (0.03), $N = 339$	-0.13*** (0.03), $N = 381$	-0.19*** (0.05), $N = 382$	-0.79*** (0.22), $N = 382$
+ Self-image control	-0.14*** (0.03), $N = 343$	-0.13*** (0.03), $N = 389$	-0.19*** (0.05), $N = 389$	-0.81*** (0.22), $N = 389$
+ Social-image control	-0.13*** (0.03), $N = 343$	-0.12*** (0.03), $N = 389$	-0.19*** (0.05), $N = 389$	-0.80*** (0.22), $N = 389$
Drop experimenter-demand resp.	-0.12*** (0.03), $N = 332$	-0.13*** (0.03), $N = 376$	-0.20*** (0.05), $N = 376$	-0.82*** (0.22), $N = 376$
Attrition: baseline carry-forward	—	—	—	—
Lee bounds [lower, upper]	[-0.18, -0.13]	[-0.15, -0.13]	[-0.19, -0.18]	—

Notes: Each cell reports the estimated treatment effect on the column outcome, its standard error in parentheses, and the sample size N. OLS / linear-probability estimates use HC3 robust standard errors; ordered- and binary-logit estimates (in log-odds) use model-based standard errors. Rows report robustness for the treatment coefficient: the main specification; then adding one baseline covariate at a time (willingness to pay for human therapy; beliefs about AI efficacy, emotional comfort, and privacy; household income); controlling for self-image and social-image concerns (Wave-5 self-reports); dropping respondents who report feeling pressured to adjust their answers; baseline carry-forward for attriters; and Lee (2009) trimming bounds. Carry-forward and Lee bounds are shown only where applicable ("—" otherwise). $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table F.10: Robustness battery: demand for medication and lay emotional support.

	Medication Week 1	Medication Week 4	Emo. support Week 4
Main specification	-0.00 (0.04), $N = 364$	-0.06* (0.04), $N = 389$	0.15 (0.19), $N = 390$
+ WTP for human therapy	-0.00 (0.04), $N = 356$	-0.07* (0.04), $N = 381$	0.11 (0.19), $N = 382$
+ AI efficacy belief	-0.00 (0.04), $N = 364$	-0.06* (0.04), $N = 389$	0.15 (0.19), $N = 390$
+ AI emotional comfort	-0.00 (0.04), $N = 364$	-0.06* (0.04), $N = 389$	0.16 (0.19), $N = 390$
+ AI privacy	-0.00 (0.04), $N = 364$	-0.06* (0.04), $N = 389$	0.15 (0.19), $N = 390$
+ Household income	-0.01 (0.04), $N = 356$	-0.07** (0.04), $N = 381$	0.14 (0.19), $N = 382$
+ Self-image control	-0.02 (0.04), $N = 360$	-0.08** (0.04), $N = 389$	0.12 (0.19), $N = 390$
+ Social-image control	-0.00 (0.04), $N = 360$	-0.05 (0.04), $N = 389$	0.09 (0.19), $N = 390$
Drop experimenter-demand resp.	-0.00 (0.04), $N = 347$	-0.06 (0.04), $N = 376$	0.24 (0.19), $N = 376$
Attrition: baseline carry-forward	—	—	—
Lee bounds [lower, upper]	[-0.03, 0.04]	[-0.06, -0.05]	[0.05, 0.11]

Notes: Each cell reports the estimated treatment effect on the column outcome, its standard error in parentheses, and the sample size N. OLS / linear-probability estimates use HC3 robust standard errors; ordered- and binary-logit estimates (in log-odds) use model-based standard errors. Rows report robustness for the treatment coefficient: the main specification; then adding one baseline covariate at a time (willingness to pay for human therapy; beliefs about AI efficacy, emotional comfort, and privacy; household income); controlling for self-image and social-image concerns (Wave-5 self-reports); dropping respondents who report feeling pressured to adjust their answers; baseline carry-forward for attriters; and Lee (2009) trimming bounds. Carry-forward and Lee bounds are shown only where applicable ("—" otherwise). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table F.11: Robustness of the treatment-by-loneliness interaction for social contact

	Week 4
Main interaction specification	1.06^{***} (0.38), $N = 390$
+ Treatment × GAD-7 (W1, std.)	1.06^{***} (0.38), $N = 390$
+ Treatment × PHQ-8 (W2, std.)	0.94^{**} (0.39), $N = 390$
+ Self-image control × Treatment	0.99^{**} (0.39), $N = 390$
+ Social-image control × Treatment	1.12^{***} (0.39), $N = 390$
Drop experimenter-demand resp.	0.98^{**} (0.38), $N = 376$

Notes: Each cell reports the coefficient on Treatment × High Loneliness (Wave 2; above-median split of the baseline UCLA-3 score, marking the scale maximum of 9) from an ordered logit of the Wave-5 frequency of meaningful social contact, its standard error in parentheses, and the sample size N. All rows include the controls of the main substitution specification. Rows add, one at a time, a competing treatment interaction with baseline symptom severity, and the Wave-5 self- and social-image controls entered both directly and interacted with treatment; the last row re-estimates the main interaction specification dropping respondents who report feeling pressured to adjust their answers. Lee (2009) bounds and baseline carry-forward are omitted because they do not apply to a subgroup-difference coefficient for an outcome without a baseline measure. Attrition is unrelated to the treatment-by-loneliness cell (interaction $p = 0.54$ in a linear probability model of outcome missingness). $^{*}\,p < 0.1$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

high by design, the CACE on the standardized Wave-4 GAD-7 score (-0.76) is only modestly larger than the ITT (-0.63).

Table F.12: Heterogeneity of the treatment effect on loneliness by baseline loneliness

	Continuous Baseline Loneliness Interaction			Median-Split Baseline Loneliness Interaction		
	UCLA Loneliness (Wave 3)	UCLA Loneliness (Wave 4)	UCLA Loneliness (Wave 5)	UCLA Loneliness (Wave 3)	UCLA Loneliness (Wave 4)	UCLA Loneliness (Wave 5)
Treatment	-0.20**	-0.41***	-0.70***	-0.27**	-0.40***	-0.72***
	(0.09)	(0.09)	(0.10)	(0.12)	(0.13)	(0.13)
UCLA Loneliness (Wave 2, std.)	0.71***	0.70***	0.71***	0.55***	0.62***	0.62***
	(0.05)	(0.06)	(0.06)	(0.08)	(0.08)	(0.08)
Treatment ×						
UCLA Loneliness (Wave 2, std.)	-0.09	-0.13	-0.12			
	(0.09)	(0.09)	(0.10)			
High Loneliness (Wave 2)				0.24	0.07	0.06
				(0.16)	(0.16)	(0.16)
Treatment ×						
High Loneliness (Wave 2)				0.17	-0.02	0.04
				(0.17)	(0.19)	(0.20)
Efficacy Human Therapy (Wave 2)	0.00	-0.00	0.00	0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Prior Therapy Experience (Wave 1)	-0.03	-0.09	-0.14	-0.03	-0.08	-0.14
	(0.09)	(0.09)	(0.10)	(0.09)	(0.10)	(0.10)
N	393	388	389	393	388	389
Adj. R^2	0.37	0.35	0.36	0.38	0.35	0.36

Notes: OLS estimates. The dependent variable is the standardized UCLA-3 loneliness score at Waves 3, 4, and 5. Columns (1)–(3) interact treatment with the standardized continuous baseline loneliness score (Wave 2); columns (4)–(6) interact treatment with an above-median indicator, which at the sample median of 8 marks participants at the scale maximum of 9. All columns use the pre-registered controls of the main loneliness specification: baseline loneliness (Wave 2), prior therapy experience (Wave 1), and beliefs about the efficacy of human therapy (Wave 2). Heteroskedasticity-robust (HC3) standard errors in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table F.13: Descriptive Wave-3 fear-of-judgment split for the Wave-5 social outcomes

	Frequency of Social Contact (Wave 5, ordered logit)	UCLA Loneliness (Wave 5, OLS)
	(1)	(2)
Treatment	0.54**	-0.60***
	(0.24)	(0.13)
Mentions Fear of Judgment or Social Avoidance (Wave 3)	-0.23	0.17
	(0.26)	(0.12)
Treatment × Fear of Judgment or Social Avoidance (Wave 3)	0.21	-0.24
	(0.36)	(0.20)
Baseline Controls	✓	✓
Obs	390	389

Notes: Column (1) reports ordered-logit coefficients for the Wave-5 frequency of meaningful social contact, with the controls of the main substitution specification; coefficients are log odds ratios, so $\exp(\beta)$ multiplies the odds of reporting a frequency above any given category. Column (2) reports OLS estimates for the standardized UCLA-3 loneliness score at Wave 5, with the controls of the main loneliness specification; heteroskedasticity-robust (HC3) standard errors. The moderator indicates whether the participant's Wave-3 open-ended explanation of their provider preference was classified (LLM classification) as mentioning fear of judgment/embarrassment or social avoidance. The classification is post-randomization and elicited differently across arms, so the split is descriptive. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table F.14: Intention-to-treat and complier average causal effect on Wave-4 anxiety

	ITT	CACE
Treatment (AI app)	-0.63***	-0.76***
	(0.08)	(0.10)
Compliance rate	83.5%	
Baseline controls	✓	✓
Obs	390	390

Notes: The dependent variable is the standardized Wave-4 GAD-7 score. The ITT column reproduces the pre-registered main specification (controlling for baseline standardized GAD-7, prior therapy experience, and baseline belief in human-therapy efficacy). The CACE column rescales the ITT by the compliance rate (the share of assigned participants completing at least one ≥10-minute AI session), which is valid under one-sided non-compliance. Robust HC3 standard errors in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

G Engagement-Based Heterogeneity in Estimated Effects

We report engagement heterogeneity with the caveat 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. Whether an individual enjoys talking to the app may be a confounder spuriously creating engagement-based heterogeneity.

We interact the assignment indicator with within-arm quartiles of cumulative therapy sessions (defined as conversations of at least five minutes) completed prior to each outcome wave.

Table G.1, 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 estimated reductions 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 responses to the intervention. 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 G.1 report these results, which indicate that these tasks may amplify the estimated benefits.

Table G.2 reports the heterogeneity by conversational turns with the AI reinforcing the pattern.

These findings of a dose–response pattern help alleviate concerns that the control groups' knowledge of their assignment causes placebo effects that can amplify estimated benefits. 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 intervention intensity per se.

Table G.1: Engagement heterogeneity

	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

Notes: OLS estimates of engagement-based 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 an indicator for assignment to the AI app, its interactions with 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 AI-app arm; the omitted interaction category is the lowest engagement group (reference). Interaction labels are: assignment × 25th–50th percentile, assignment × 50th–75th percentile, and assignment × 75th–100th percentile (when identified by the data). Heteroskedasticity-robust (HC3) standard errors are used. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table G.2: GAD-7 and PHQ-8 on total conversational turns

	GAD-7 (Wave 4)	PHQ-8 (Wave 4)
Total AI Turns (Weeks 1-4, per 100)	-0.06*	-0.13***
	(0.03)	(0.03)
GAD-7 Score (Wave 1)	0.64***	
	(0.15)	
PHQ-8 Score (Wave 2)		0.60***
		(0.08)
N	190	190
Adj. R^2	0.11	0.25

Notes: OLS estimates of Wave-4 GAD-7 and PHQ-8 scores on total AI conversational turns accumulated over the first four weeks of app use (sum of AI and human turns across Weeks 1–4, rescaled per 100 turns), restricted to participants assigned to the AI app. Each regression controls for the corresponding baseline score (GAD-7, Wave 1; PHQ-8, Wave 2), therapy history (Wave 1), and perceived efficacy of human therapy (Wave 2). Heteroskedasticity-robust (HC3) standard errors are shown in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

H The Trial Sample Compared to a Survey Representative in Gender, Age, and Ethnicity

Our sample was recruited through social media advertisements promoting an AI-based mental health tool, a channel that selects for participants with relatively low concerns about sharing personal information with AI and high GAD-7 scores: early adopters of AI therapy with high need for support. To compare our trial sample to the more general population, we conduct a survey with $N = 300$ respondents on Prolific that is representative of the US population in terms of gender, age, and ethnicity. Table H.1 compares the trial control arm to survey participants split by whether they are interested in AI psychotherapy, and Figure H.1 compares their beliefs about and preferences over AI versus human therapy. Relative to the survey, the trial control arm is younger, more symptomatic, more treatment-experienced, and markedly more favorable toward AI across effectiveness beliefs, stated preferences, and perceived advantages. Among survey respondents, those interested in AI therapy differ from those not interested mainly in symptom burden and comfort disclosing to AI and less in what they say they want from therapy (panel D of Figure H.1), suggesting that the currently uninterested may be potential future adopters rather than a fundamentally distinct population.

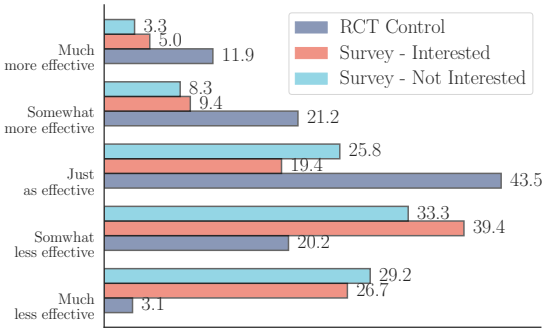
Table H.1: Baseline descriptives for the RCT control group compared to survey participants

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)

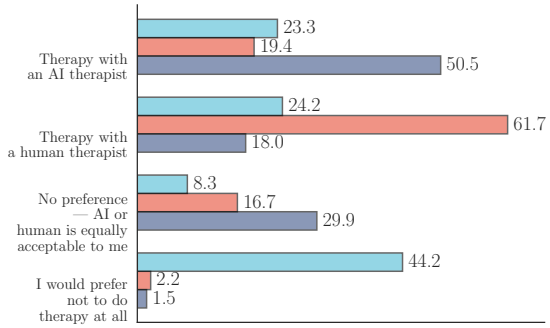
Notes: The first column contains baseline characteristics for RCT participants assigned to the control group. The second and third columns report the same variables for survey participants split by whether they report being interested in engaging with AI psychotherapy.

Figure H.1: Therapy preferences in RCT and survey

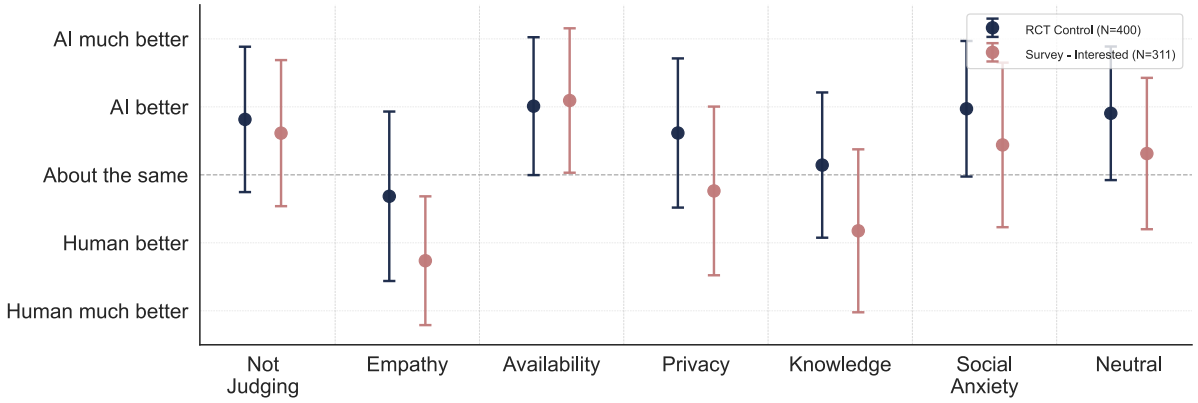
(a) Belief about effectiveness of AI vs. human therapist (% of responses)



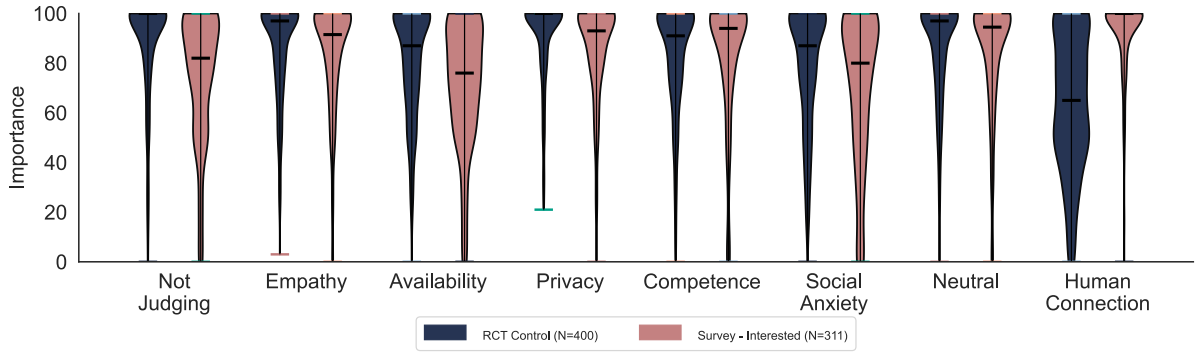
(b) Preference for AI vs. human therapy (% of responses)



(c) Assignment of therapy attributes as qualities of AI vs. human



(d) Self-reported importance of therapy attributes



Note: (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) Points show item means on a five-point scale recoded 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 AI-app arm (blue), RCT control arm (red), and survey respondents (cyan). Characteristics were identified from open-ended Wave-1 responses via LLM classification (Section D) and assessed quantitatively in Wave 4. Black lines indicate medians; violin width represents response density. $N = 400$ (RCT); $N = 300$ (survey).

I Additional Sample Descriptives

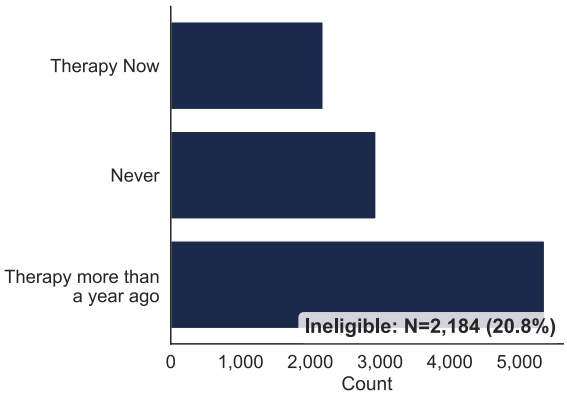
We recruited participants via targeted advertisements on Facebook and Instagram (Meta Platforms), using the Meta Pixel to track study-website visits and to optimize advertising delivery. Following screening, we transmitted an eligibility indicator back to Meta to preferentially allocate impressions toward users likely to meet inclusion criteria. We discuss the selection process in Section I.1 and then describe the final randomized sample in Section I.2.

I.1 Sample Selection

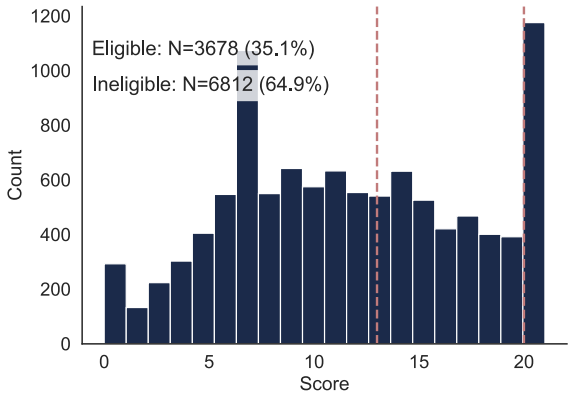
The pool of baseline respondents is pre-selected by construction: individuals must first be reached by platform targeting and then opt in by clicking the advertisement. Within this selected pool, we impose four additional eligibility criteria. Two follow directly from the trial's clinical focus on CBT for anxiety: we require a baseline GAD-7 score between 13 and 20 (35.1% of Wave 1 respondents fall outside this range), and we exclude individuals currently in therapy to avoid spillovers with existing care (Figure I.1). To mitigate non-compliance, we further restrict on two variables predictive of adoption of AI-based tools: comfort engaging with an AI system for emotional support and perceived privacy risk, benchmarked against the survey described in Section H, representative of US adults in gender, age, and ethnicity (Figure I.2). In the representative sample, 48.3% report comfort < 50, compared with 24.4% in Wave 1; 39.7% report being very or extremely concerned about privacy, versus 26.8% in Wave 1. These comparisons indicate that our sample is selected in predictable ways but still maps onto a substantial segment of the general population that would plausibly adopt AI-based therapy.

Figure I.1: Eligibility screener: Wave 1 full sample

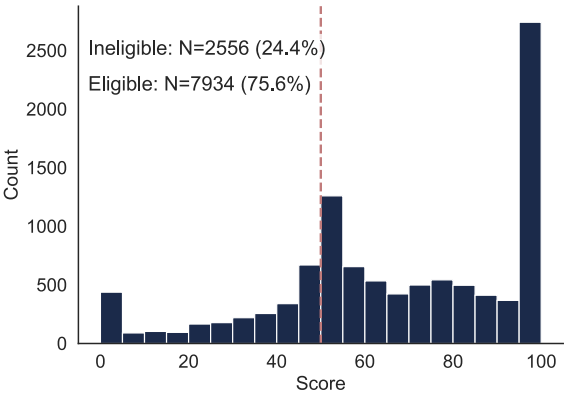
(a) Therapy history (Wave 1).



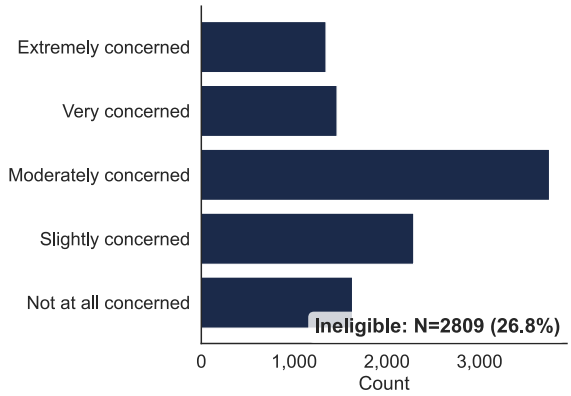
(b) GAD-7 score (Wave 1).



(c) AI emotional comfort (Wave 1).



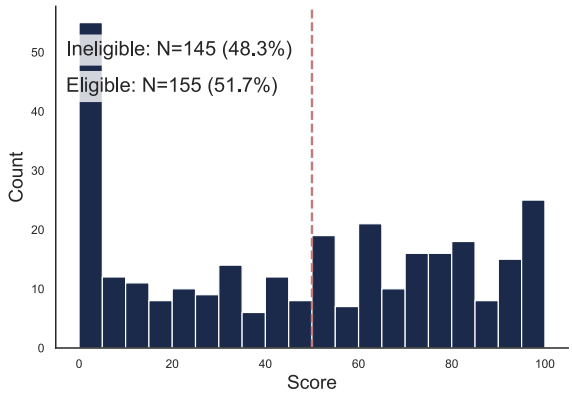
(d) Privacy concerns about AI (Wave 1).



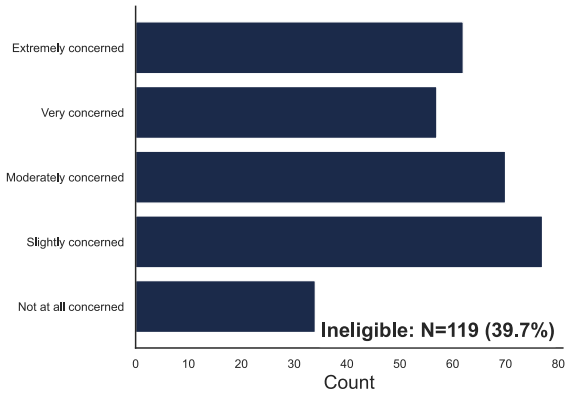
Note: This figure summarizes the Wave 1 eligibility criteria across the full sample, showing distributions for therapy history, AI privacy concern, AI emotional comfort, and GAD-7 scores.

Figure I.2: AI attitudes: representative survey

(a) AI emotional comfort.



(b) Privacy concerns about AI.



Note: This figure summarizes AI-related eligibility criteria from the representative survey, showing distributions for AI emotional comfort and privacy concerns about AI.

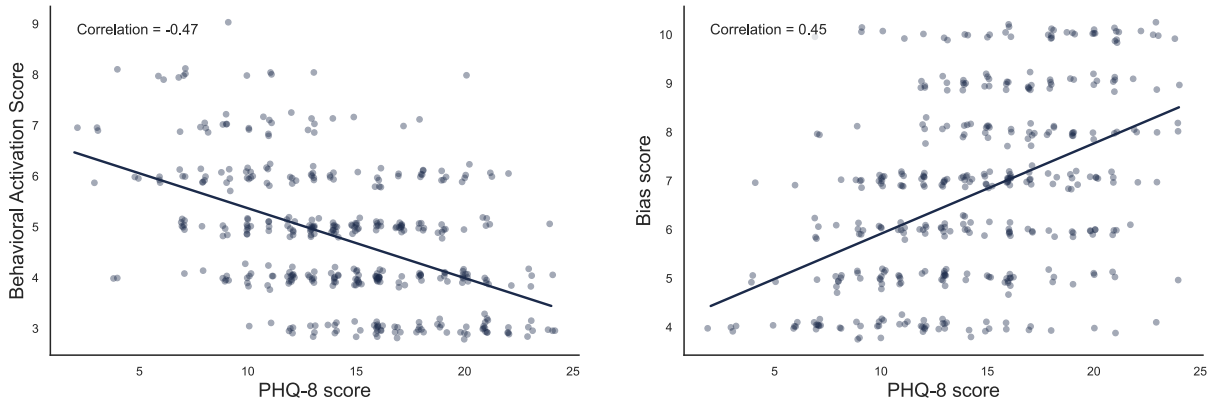
I.2 Randomized Sample Descriptives

In line with the goal of this study and the pre-registered sample selection, the sample represents a population high in symptom burden and 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 $\geq$ 15). Most also report elevated depressive symptoms (PHQ-8). As expected, higher PHQ-8 scores are associated with lower behavioral activation and a stronger negativity bias at baseline (Figures I.3a and I.3b).

Figure I.3: PHQ-8 relationships with behavioral activation and negativity bias

(a) PHQ-8 and behavioral activation scores (Wave 2). (b) PHQ-8 and negativity bias scores (Wave 2).



Note: Scatter plots with small random jitter for visualization and an OLS regression line.

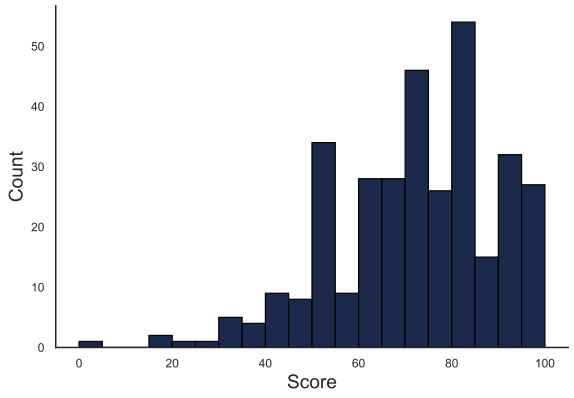
Attitudes towards Psychotherapy Approximately 80% of the sample have previously engaged in psychotherapy and report prior use of psychiatric medication. Despite this, respondents express substantial self- and public-stigma related to mental health, measured using the instrument from Roth et al. (2024).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

Baseline Beliefs about AI Because our recruitment advertisement described the study as evaluating an AI-based support tool, participants are positively predisposed toward AI. As shown in Figure I.4,

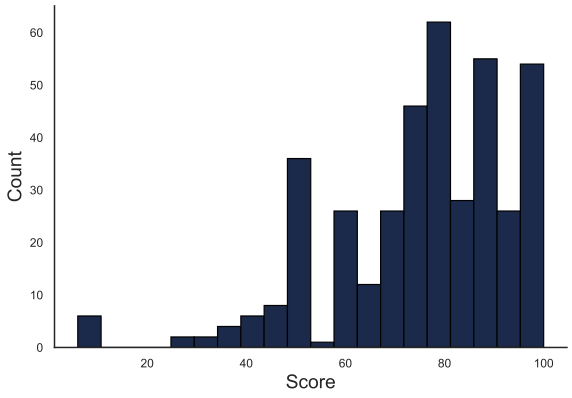
most respondents report frequent use of large language model tools, rate the accuracy of AI-generated medical information as relatively high, and tend to believe that AI-delivered therapy could be effective for them personally.

Figure I.4: Baseline descriptives on beliefs about AI in the full randomized sample ($N = 400$)

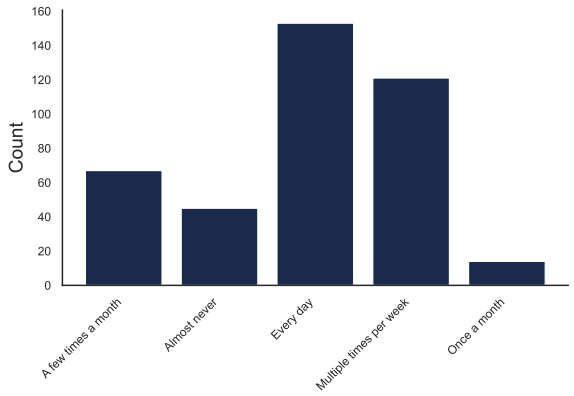
(a) Effectiveness beliefs about AI therapy



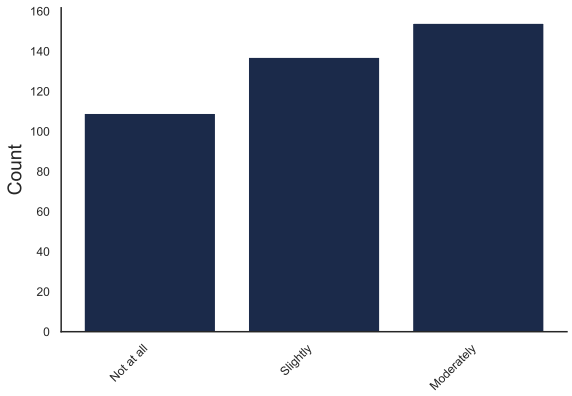
(b) Estimated AI medical advice accuracy



(c) AI usage frequency



(d) Privacy concerns about AI



Note: 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 of AI-delivered psychotherapy. Perceived AI medical accuracy was assessed with an estimate item from 0 to 100, where higher values indicate greater perceived alignment of AI outputs with scientific consensus. AI usage frequency and the level of privacy concern 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 I of the main text, participants assigned to the two arms 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 estimates with these additional covariates and show that our estimates remain robust.

J 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 individuals assigned to the information arm 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:

$$Y_i = \beta_0 + \beta_1 Info_i + \beta_2 (Prior_i - 92.6) + \varepsilon_i, \tag{1}$$

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 J.1 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 J.1: 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 the effect of the information intervention 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 intervention 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.

K 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:

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

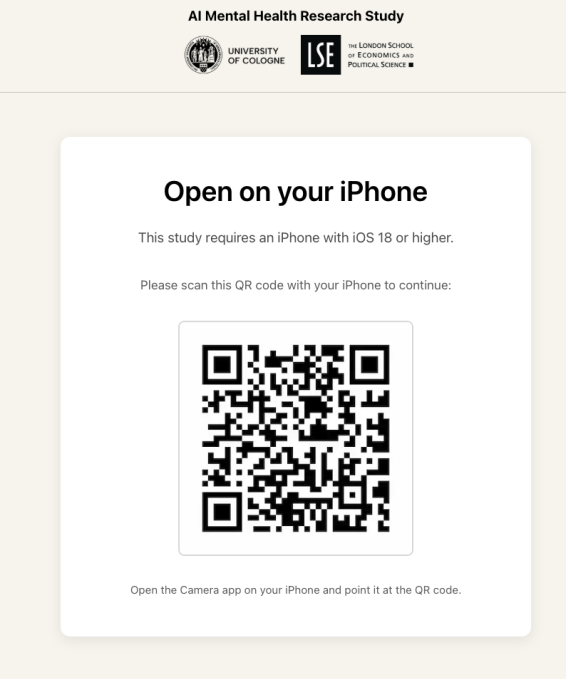
where Y_i equals 1 if the participant indicates willingness to participate, $Info_i$ is an indicator of whether the person has seen the actual information or not, and $Prior_i$ is the prior belief of person i. Primary parameter of interest is β_1; any other potential analyses of heterogeneity and using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

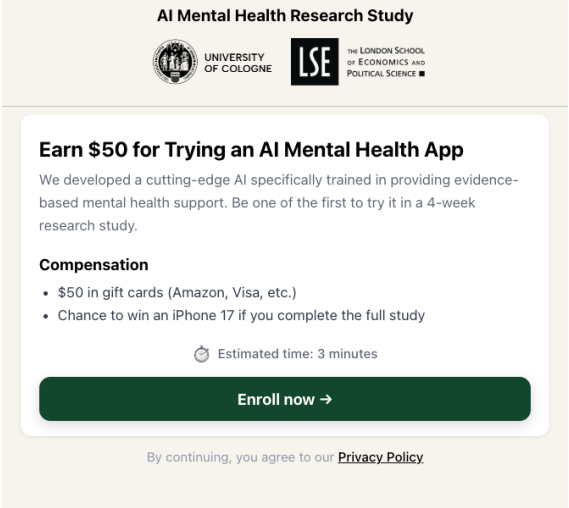
The relevant sample for this analysis comprises all participants who get to the trial information page and click on either the button that they are willing to participate or that they are not willing to participate.

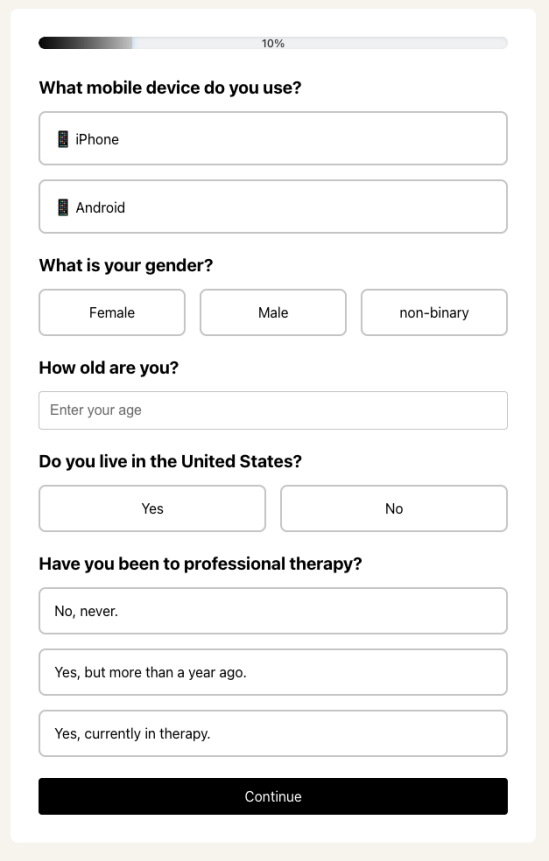
- Mechanisms of the app (automatic thoughts, behavioral activation, cognitive biases) all measured by scales elicited in all waves
- We also test heterogeneity of treatment effects by mechanisms, cross-randomization, and engagement, although we acknowledge that we might be underpowered.
- Estimation 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.

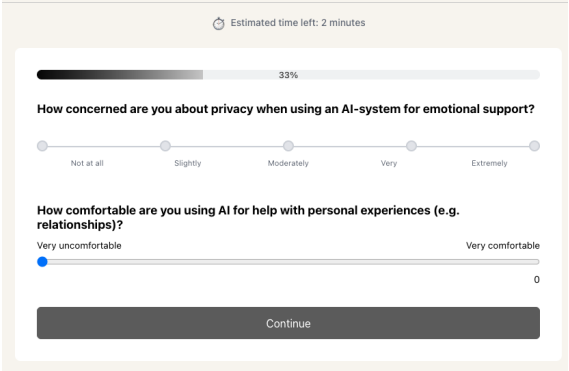
L Survey Instructions

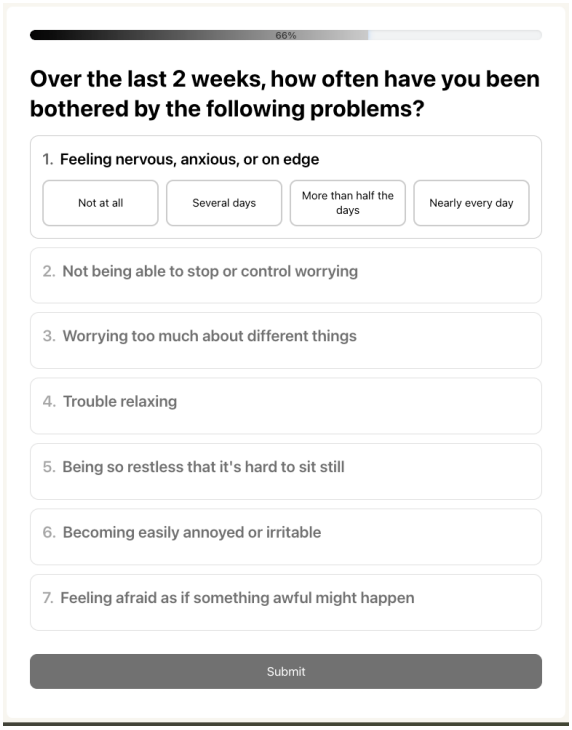
L.1 RCT Survey Instructions Wave 1

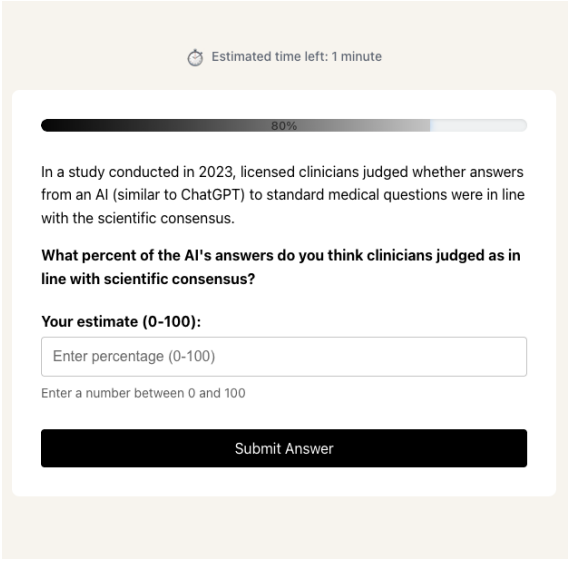


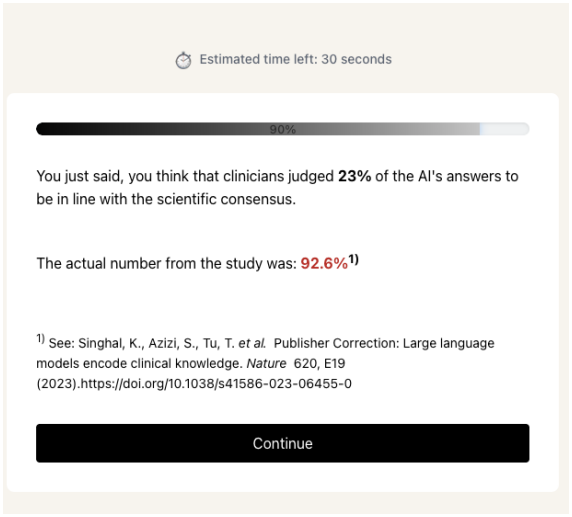


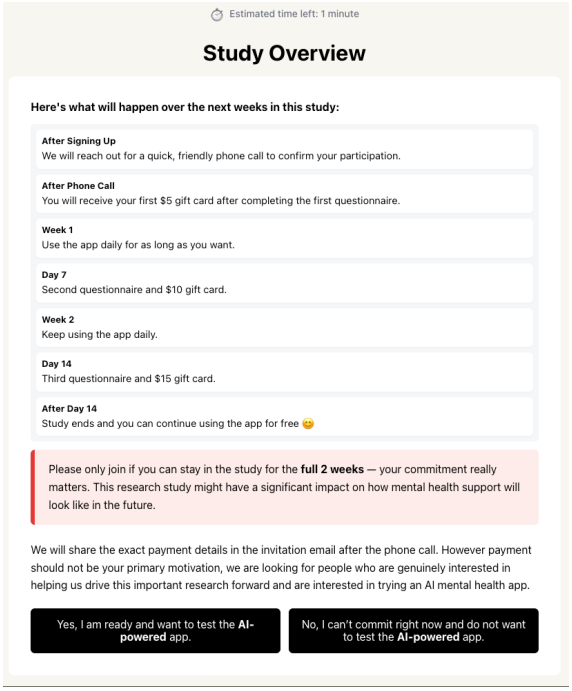





66%
Over the last 2 weeks, how often have you been bothered by the following problems?
1. Feeling nervous, anxious, or on edge
Not at all
Several days
More than half the days
Nearly every day
2. Not being able to stop or control worrying
3. Worrying too much about different things
4. Trouble relaxing
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.

L.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: [redacted]

Test an AI Powered Mental Health App

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

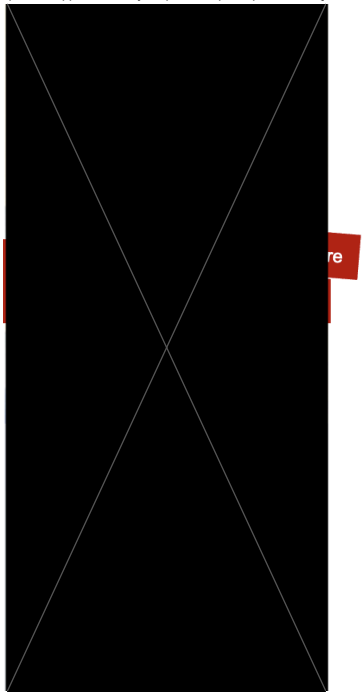
Follow these three steps **now:**

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!

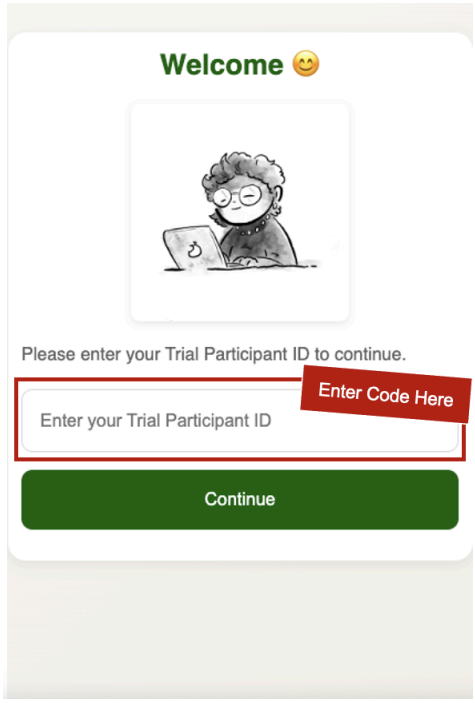
To access the resources you can use during the waiting period, follow these three steps:

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 ✉ 😊

Your Bonus Payments for Using ✉

- **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 ✉ 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

L.3 RCT Survey Instructions Wave 3

Consent ✉ **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** a [redacted]	
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 [redacted]

How much would you willing to pay for a 50-minute [redacted] session?

Amount in $

How satisfied were you with your experience using the [redacted] app?

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

How likely are you to recommend the [redacted] 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 [redacted] 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

L.4 RCT Survey Instructions Wave 4

Consent [redacted] 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 [redacted] new

Do you feel [redacted] is a safe system **compared to** traditional AI chatbots?

- ○ Yes
- ○ No

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

[ ]

therapeutic_alliance

You have been using the [redacted] 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 [redacted] 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 [redacted] 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 [redacted] 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

L.5 RCT Survey Instructions Wave 5

Consent_[redacted] 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_[redacted]

Did your belief change about what AI-therapy actually is after using the [redacted] 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 [redacted] app
- No, my beliefs stayed the same

Did using the [redacted] 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

L.6 Representative Survey Instructions

Consent_[redacted] 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 **<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
[ ]

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

Bursztyn, Leonardo, Ingar K Haaland, Nicolas Röver, and Christopher Roth, "The Social Desirability Atlas," Technical Report, National Bureau of Economic Research 2025.

Chopra, Felix and Ingar Haaland, "Conducting qualitative interviews with AI," *CESifo Working Paper No. 10666*, 2023.

Grootendorst, Maarten, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

Harrer, Mathias, Antonia A. Sprenger, Paula Kuper, Eirini Karyotaki, and Pim Cuijpers, *metapsyData: Access the Meta-Analytic Psychotherapy Databases in R* Metapsy, Metapsy, 2022.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76 (2009), 1071–1102.

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

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

Röver, Nicolas, Lukas Wolf, Valerie Forman-Hoffman, Patricia Areán, and Bennet Lucca Feld, "Short-term Evaluation of a Voice-Based Generative AI Emotional Support Companion: A Randomized Controlled Trial," *PsyArXiv preprint*, July 2026.

Singhal, Karan, 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, "Large language models encode clinical knowledge," *Nature*, 620 (2023), 172–180.