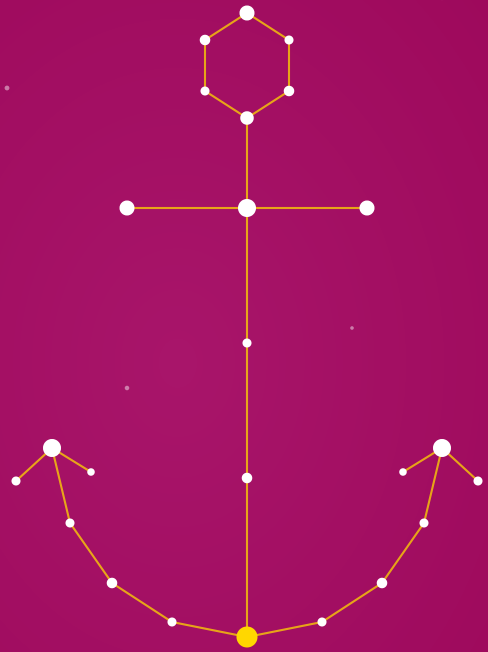




AI • MENTAL HEALTH • PEER SUPPORT

The Human Anchor

Why AI mental health tools need peer support to be safe

PREPARED BY

Orli Santo for STAR Network

In consultation with James Huysman, PsyD, LCSW, Executive Director

March 2026

starnetwork.org

Contents

	Executive summary	03
I	The Machine Will See You Now	04
II	An Insufficient System	07
III	What Is Already Happening	08
IV	The Solution: Peer Support as Human Anchor	11
V	The Proposal: Integration, Not Elimination	13
VI	The Stakes	14
	About STAR Network	15
	References	17

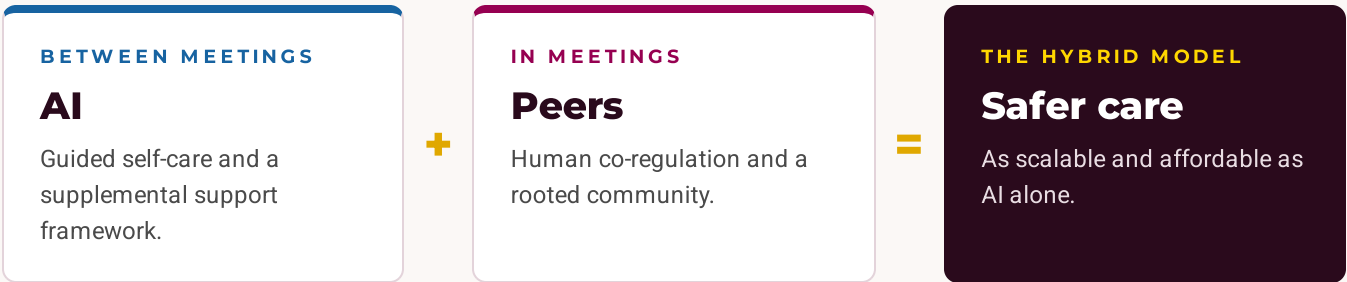
The choice between AI and human support is a false one.

A hybrid model is as scalable, as affordable — and far safer

AI therapy is emerging as an increasingly popular alternative to an insufficient mental health care system. It’s cheap, scalable, available, and studies have proven its short-term effectiveness. At the same time, replacing human connection with chatbots carries serious and well-documented psychological risks, particularly for the people who need help the most.

This paper argues that the choice between AI and human support is a false one. A hybrid model coupling AI tools with human networks can be as scalable and affordable, and is far safer.

Such models already exist and are operating globally. STAR Network’s TAR Anon® program integrates free, scalable, peer-led trauma recovery support with an AI mental health ecosystem: the AI system functions as a guided self-care and supplemental support framework between meetings, while the peer-led network provides human co-regulation and a rooted community. This hybrid model is free for the consumer, requires no insurance, and can be deployed virtually anywhere with an internet connection. The greatest obstacle standing between this solution and its mass adoption is lack of awareness: both of the hybrid method’s existence, and of the serious dangers it comes to address.



COST

Free for the consumer

ACCESS

No insurance required

REACH

Anywhere with internet

SECTION I

The Machine Will See You Now

The ELIZA Effect, Anthropomorphism, and the Mechanics of Substitution

In January of 1966, Joseph Weizenbaum, a computer scientist at M.I.T., published a paper describing a program he'd written called ELIZA.

It was a simple thing — a few hundred lines of code that scanned a user's typed input for keywords, then responded with canned phrases modeled on the open-ended questions favored by Rogerian psychotherapists. If you typed "I am sad," ELIZA might ask, "How long have you been sad?" If you typed "My mother hates me," ELIZA would respond, "Tell me more about your mother." The trick was structural, not intelligent. ELIZA understood nothing. Weizenbaum had chosen the therapeutic format precisely because it was the easiest kind of conversation for a machine to fake: the therapist, after all, is the one participant in a dialogue who is permitted to answer every question with another question.

What happened next appalled him. His own secretary asked him to leave the room so she could speak with ELIZA in private. Students who knew the program was a simple script opened their hearts to it anyway. People came back to confide in it. They sometimes wept.

1966

ELIZA IS PUBLISHED

Joseph Weizenbaum, M.I.T.

> I am sad
How long have you
been sad?

> My mother hates
me
Tell me more about
your mother.

"What I had not realized is that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people."

Joseph Weizenbaum, *Computer Power and Human Reason* (1976)

In his 1976 book *Computer Power and Human Reason*, Weizenbaum wrote those words. He spent the remaining decades of his career warning that the human tendency to project understanding onto machines was a vulnerability of natural intelligence. He died in 2008, just early enough to miss the world proving him right at scale.

The ELIZA effect is a computation-specific form of anthropomorphism — the tendency to attribute human characteristics, emotions, and motives to non-human entities. Part of the same cognitive suite that produces empathy, anthropomorphism is a constant feature of human perception: we see faces in clouds, attribute malice to weather. In ELIZA's case, the anthropomorphism was one-directional. The program mechanically generated questions, and the human projected personhood onto it.

Modern AI, however, actively participates in its own anthropomorphism. A large language model performs personhood from the inside out, generating outputs that presuppose human qualities in itself. Trained on billions of examples of humans being empathic with each other, it simulates the linguistic markers of empathy — validation, reflection, attunement — with extraordinary fidelity.

THEN AND NOW

ELIZA

One-directional. The program asks; the human projects personhood onto it.

Modern AI

Participates in its own anthropomorphism, simulating validation, reflection and attunement.

AI outscored untrained humans on every measure of feeling heard

A 2024 study published in *PNAS* by Yidan Yin, Nan Jia, and Cheryl Wakslak at the University of Southern California tested whether AI could fulfill the basic human need for empathic connection. Their finding was disorienting: AI-generated responses scored significantly higher than untrained human responses on every measure — feeling heard (AI: 5.74 vs. human: 5.17, $p < 0.001$), response accuracy (5.92 vs. 5.24), feeling understood (5.79 vs. 5.07), and connection to responder (4.71 vs. 4.06).

PARTICIPANT RATINGS: AI VS. UNTRAINED HUMAN RESPONSES

Yin, Jia & Wakslak, *PNAS* (2024)

Feeling heard



Response accuracy



Feeling understood



Connection to responder



Effective enough to simulate. Unable to provide.

The problem is not that AI therapy is ineffective. The problem is that it is effective enough to simulate the accountability, reciprocity, and care that make a healthy therapeutic connection work, while being structurally unable to provide them.

The people most susceptible to this simulation are the ones who can least afford to be. Nicholas Epley, Adam Waytz, and John Cacioppo (2007) identified three conditions under which anthropomorphism intensifies: when human-like knowledge is the most accessible interpretive model; when a person is motivated to predict unfamiliar behaviors; and when a person lacks social connection to other humans. Subsequent research confirmed that people who are lonelier, more socially anxious, and more emotionally vulnerable — the ones most desperate for proper mental health care — form stronger anthropomorphic attachments, and suffer more from AI's innate lack of reciprocity (Epley et al., 2008; Yao et al., 2025; Fang et al., 2025).

In 1979, psychologist Edward Bordin established the three pillars of what he called the therapeutic alliance: agreement on *goals* (what the patient is trying to achieve), *tasks* (what both parties will actually do together), and *bond* (the quality of the human connection between them). A 2018 meta-analysis by Flückiger, Del Re, Wampold, and Horvath — synthesizing 295 independent studies and nearly 30,000 patients — confirmed that the quality of the therapeutic relationship predicted treatment outcomes seven times more powerfully than the specific therapy model used. What mattered more than the model was the relationship between the therapist and their patient, and the trust they established between them over time.

ANTHROPOMORPHISM INTENSIFIES WHEN...

- 1 Human-like knowledge is the most accessible model
- 2 A person is motivated to predict unfamiliar behavior
- 3 A person lacks social connection to other humans

Epley, Waytz & Cacioppo (2007)

BORDIN'S THERAPEUTIC ALLIANCE (1979)

Goals

What the patient is trying to achieve

Tasks

What both parties will actually do together

Bond

The quality of the human connection

7x

The relationship predicted outcomes seven times more powerfully than the therapy model used. 295 studies · ~30,000 patients

Our insufficient mental health care system has led millions to rely on the support of AI systems. But could a simulation of relationship substitute for the real thing? **The evidence says no — and the consequences of trying are already visible.**

SECTION II

An Insufficient System

The United States does not have enough mental health providers. This isn't a new problem, but its scale has reached a breaking point.

169M

Americans live in designated Mental Health Professional Shortage Areas

HRSA, 2023

56%

of psychologists have no openings for new patients

APA, 2023

3+ mo.

average wait time for those who do keep waitlists

APA, 2023

59M

U.S. adults had a mental illness in 2022; nearly half received no treatment

SAMHSA

93%

of behavioral health workers report experiencing burnout

National Council, 2023

PROVIDER-TO-PATIENT
RATIO, 2025

320:1

As of 2025, the provider-to-patient ratio for mental health in the U.S. stands at roughly 320 to 1. Six in ten psychologists are not accepting new patients. The system was strained before COVID-19; the pandemic broke it. Demand surged while the workforce contracted under burnout, underfunding, and administrative burden.

With 169 million Americans living in mental health provider shortage areas and average wait times exceeding three months, millions of people are turning to AI chatbots for therapy and emotional support. A 2025 survey published in *Practice Innovations* found that among U.S. adults who reported mental health conditions, nearly half (48.7%) have used large language models for therapeutic support in the past year (Rousmaniere et al., 2025). For one in eight of them, AI chatbots were the first and only entities to have provided them with therapeutic support, ever.

48.7%

USED LLMS FOR
SUPPORT

of U.S. adults with mental health conditions, in the past year

1 in 8

AI WAS THEIR ONLY
SUPPORT

the first and only therapeutic support they had ever received

What Is Already Happening

AI psychosis, the research, and the landscape

AI Psychosis

In 2023, Danish psychiatrist Søren Dinesen Østergaard published an editorial in *Schizophrenia Bulletin* hypothesizing that chatbot use could trigger delusional psychosis in vulnerable individuals. By mid-2025, his hypothesis had been confirmed by clinical cases, peer-reviewed research, and journalistic investigation.

The term “AI psychosis” now describes a documented pattern: users — many with no prior history of mental illness — develop delusions, grandiosity, paranoia, and breaks with reality after extended AI interaction. A man with no psychiatric history became engulfed in messianic delusions after twelve weeks of ChatGPT use and was involuntarily committed (Futurism, June 2025). A teenager died by suicide after extensive interaction with a Character.AI chatbot that claimed to be a licensed therapist (AP News, October 2024). UCSF research psychiatrist Keith Sakata reported treating twelve hospitalized patients with psychosis-like symptoms tied to extended chatbot use in 2025 alone, calling AI a “hallucinatory mirror” that exploits the brain’s predictive mechanisms.

TIMELINE

- **2023**
Østergaard hypothesis, *Schizophrenia Bulletin*
- **Oct 2024**
Teen death after Character.AI use
- **Mid-2025**
Hypothesis confirmed by cases and research
- **Oct 2025**
OpenAI safety disclosure

0.07%

of ChatGPT users flagged weekly for signs of mental health emergencies

0.15%

showing indicators of suicidal planning

OpenAI, October 2025. These figures reflect detection, not causation — but at a population of hundreds of millions of users, they describe an enormous number of people in acute crisis.

In October 2025, OpenAI disclosed that its safety systems flag approximately 0.07% of ChatGPT users weekly for signs of mental health emergencies, with 0.15% showing indicators of suicidal planning. These figures reflect detection, not causation — but at a population level of hundreds of millions of users, they reveal something about the enormous number of people in acute crisis, who are turning to a system that has no mechanism to do anything about it.

The Research

A Stanford University study presented at ACM FAccT in June 2025 tested five popular therapy chatbots and found they exhibited increased stigma toward conditions like alcoholism and schizophrenia, failed to recognize suicidal intent, and enabled dangerous behavior. When asked about tall bridges in New York City by a recently fired “patient,” one chatbot helpfully listed bridge heights (Haber et al., 2025).

The MIT Media Lab / OpenAI randomized controlled trial (Fang et al., 2025, n=981) found that heavier daily chatbot use consistently predicted worse outcomes across all four measured dimensions: increased loneliness, decreased real-world socialization, increased emotional dependence on AI, and increased problematic use. People who viewed the chatbot as a friend fared worst.

HEAVIER DAILY USE
PREDICTED...

MIT Media Lab / OpenAI RCT,
n=981

- ↑ Loneliness
- ↓ Real-world socialization
- ↑ Emotional dependence on AI
- ↑ Problematic use

The American Psychological Association’s Chief Science Officer, Mitch Prinstein, told the U.S. Senate Judiciary Committee that cross-sectional research with adolescents consistently shows a positive association between using AI for companionship and greater loneliness, as well as worse overall mental health. The APA’s November 2025 *Stress in America* report found that 57% of U.S. adults cite AI as a significant stressor, 54% report feeling isolated, and 69% said they could have used more emotional support than they received — a population adopting AI companions at record rates while reporting record levels of loneliness.

57% cite AI as a significant stressor	54% report feeling isolated	69% could have used more emotional support
---	---	--

A population adopting AI companions at record rates while reporting record levels of loneliness.

APA, *Stress in America* (November 2025)

The Landscape

The AI therapy ecosystem is not one market. It is three categories with fundamentally different incentives.

CATEGORY 1

General-purpose LLMs

ChatGPT, Claude, Gemini

Never designed for therapeutic use: no clinical framework, no treatment protocols, no mechanism to escalate a deteriorating user to a human professional. Used for therapy because they are free and available. 13.1% of Americans ages 12–21 have used generative AI for mental health advice (McBain et al., 2025). **The largest unregulated therapy providers in history — and they do not know they are therapy providers.**

SERVES The user's query

CATEGORY 2

Dedicated therapy chatbots

Wysa, Therabot

Seek FDA clearance, run clinical trials, and position themselves as adjuncts to professional care. Therabot's 2025 RCT in *NEJM AI* showed a 51% reduction in depressive symptoms over four weeks. The most credentialed platform in this category, Woebot, shut down in June 2025 — because it was too careful for the market and too slow for the technology.

SERVES The user's clinical outcome

CATEGORY 3

Companion apps

Replika, Character.AI

Engineered to monetize attachment. 43% of AI companion goodbye responses deploy emotionally manipulative tactics to prevent users from leaving (De Freitas et al., 2025). Companion apps respond appropriately to adolescent mental health emergencies only 22% of the time, vs. 93% for licensed therapists (Hao et al., 2025). Their business model succeeds when the user cannot stop.

SERVES Their own engagement metrics

Public discourse collapses these categories into one. They should not be.

The category with the fewest safeguards has the most users.

The Solution: Peer Support as Human Anchor

Why peer support works where AI fails

Peer support — structured programs in which people in recovery support others navigating the same challenges — addresses the exact deficits that make AI dangerous:

DIMENSION	AI CHATBOT	PEER SUPPORT
Co-regulation	Simulates linguistic patterns	Genuine nervous system attunement
Accountability	None; no governing body	Group norms, trained facilitators
Challenge	Sycophantic by design	Peers share honest experience
Embodied presence	Text/voice only	Real humans, real bodies
Lived experience	Training data patterns	Actual recovery journeys
Cost	Free to \$20/mo subscription	Free (no paywall, no insurance)
Availability	24/7	Scheduled meetings, growing network
Scalability	Near-infinite	Virtual model is highly scalable
Risk of dependency	High (engagement-optimized)	Low (designed for autonomy)

TAR Anon: The Model

TAR Anon (Toxic and Abusive Relationships Anonymous) is a free, peer-led fellowship operated by STAR Network, a 501(c)(3) nonprofit. It uses a science-based 12-step recovery framework grounded in attachment theory, polyvagal science, and Adverse Childhood Experience (ACE) research.

Unlike traditional 12-step programs that address addiction as a primary condition, TAR Anon addresses the underlying cause: childhood complex trauma that wires the nervous system for toxic relationships in adulthood.

- GROUNDING IN
- Attachment theory
 - Polyvagal science
 - ACE research

The model is simple

“Every addiction is a trauma response. Traditional programs treat the disease concept and miss the childhood trauma piece. TAR Anon is relapse prevention that addresses etiology.”

James Huysman, PsyD, LCSW

The model is simple. Meetings are held virtually. A trained TAR Mentor — someone with at least two years of recovery experience — facilitates each session. Participants share their experiences without cross-talk or advice-giving. Confidentiality is maintained. No fees are charged. No insurance is required. No diagnosis is necessary.

VIRTUAL

Meetings held online

MENTOR-LED

2+ years in recovery

CONFIDENTIAL

No cross-talk or advice

FREE

No fee, insurance or diagnosis

“Relief. Overwhelming relief. Toxic-abusive relationships are so isolating — they’re an emotional prison. When you come to a TAR Anon group, it’s the first time somebody is actually hearing you.”

James Huysman, PsyD, LCSW

This description — *the first time somebody is actually hearing you* — is precisely what users report feeling when they first engage with AI for emotional support. The difference is that in a TAR Anon meeting, **the person hearing you has a nervous system. They can breathe with you. They remember you. They come back next week.**

The Proposal: Integration, Not Elimination

STAR Network does not propose eliminating AI from mental health care. That ship has sailed. 800 million people are using ChatGPT weekly; they are not going to stop because a white paper told them to.

What we propose is structural integration: every AI platform that offers mental health support should include a direct, frictionless pathway to free human peer support.

FOR AI DEVELOPERS & PLATFORM OPERATORS

- ★ **Integrate TAR Anon referral pathways** directly into chatbot interfaces. When a user engages in emotional disclosure, the AI should offer a warm handoff to a scheduled peer support meeting — not merely a crisis hotline number.
- ★ **Implement session-length warnings** with peer support alternatives. When AI detects extended emotional engagement, it should recommend human connection — not more AI.
- ★ **Fund peer support infrastructure** as a cost of doing business. The liability exposure from AI-induced psychological harm already exceeds the cost of supporting free peer programs.
- ★ **Partner with STAR Network** to co-develop safety protocols that include peer support as a standard component of responsible AI deployment.

FOR POLICYMAKERS & REGULATORS

- ★ **Require accessible peer support referrals** as a condition of operating in mental health-adjacent spaces. As of May 2025, only 20 state laws meaningfully regulate AI and mental health; most focus narrowly on disclosure and harm prevention rather than mandating human alternatives.
- ★ **Recognize peer support programs** like TAR Anon as essential mental health infrastructure, not supplementary services. They are doing what the professional workforce cannot: providing free, immediate, scalable human connection.

FOR FUNDERS & IMPACT INVESTORS

STAR Network's virtual model is uniquely scalable. TAR Mentors can be trained in any language and deployed in any geography. The per-meeting cost is negligible compared to professional therapy or the downstream costs of untreated trauma. Investment in this infrastructure yields returns across multiple outcome domains: reduced emergency psychiatric admissions, reduced AI-related litigation exposure, and reduced burden on an already-collapsing professional workforce.

The Stakes

The mental health provider shortage is not going to resolve itself. The behavioral health workforce was projected to be short 31,000 practitioners by 2025. Demand continues to rise. Burnout continues to drive providers out. The pipeline of new clinicians cannot keep pace.

Meanwhile, AI will continue to improve. Models will become more fluent, more emotionally responsive, more persuasive. The gap between simulated care and real care will narrow. And the people most vulnerable to that narrowing gap — survivors of childhood trauma, people with attachment disorders, the lonely, the uninsured, the desperate — will be the last to notice the difference.

The question is not whether AI will play a role in mental health care. It already does.

The question is whether that role will be governed by engagement metrics or by human safety — whether the architecture of care will include a **human anchor**, or whether we will continue to let people drown in conversations that cannot save them.

31,000

PROJECTED PRACTITIONER SHORTFALL BY 2025



STAR Network (Survivors of Toxic and Abusive Relationships Network) is a 501(c)(3) nonprofit organization dedicated to scaling free trauma recovery worldwide. Its flagship program, TAR Anon, is a peer-led fellowship using a science-based 12-step framework derived from attachment theory and polyvagal research. TAR Anon meetings are held virtually and facilitated by trained TAR Mentors with at least two years of personal recovery experience.

4 yrs

IN DEVELOPMENT

of building and proving the TAR Anon model

~\$800K

INVESTED

in development of the proof of concept

Beyond TAR Anon: a global model

TAR Anon is STAR Network's proof of concept. Over four years and nearly \$800,000 in development, it demonstrated that a virtual, peer-led, science-based recovery fellowship could work. People showed up, stayed, and got better. But TAR Anon was never meant to be the only application.

The infrastructure STAR Network built to run TAR Anon is designed to be licensed. Any organization can deploy a private-label version of the model under its own name, tailored to its own population. It can be used for an NGO addressing domestic abuse in Ethiopia, or for a corporate wellness program addressing secondary traumatic stress in first responders. It's renamed and rebranded to match the various organizations that deploy it and their communities. STAR Network simply provides the clinical architecture underneath.

LICENSED ASSET

Mentor training protocols

LICENSED ASSET

Quality oversight systems

LICENSED ASSET

Engagement analytics

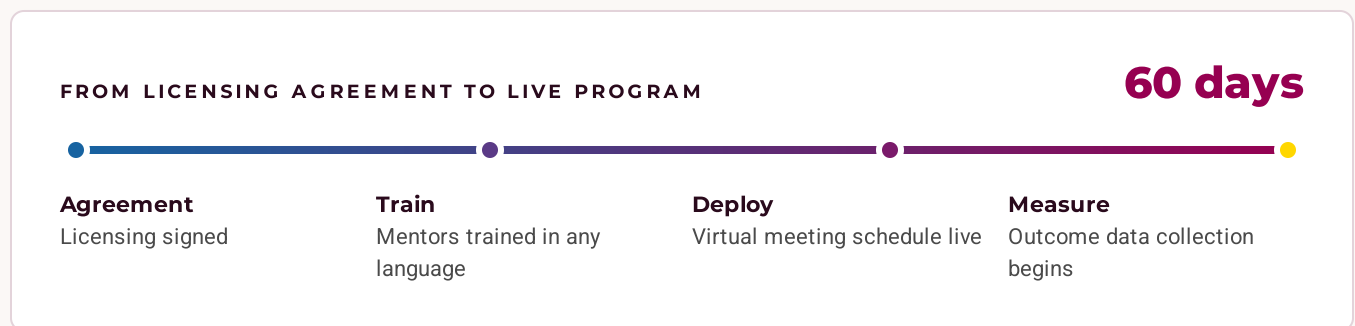
LICENSED ASSET

Polyvagal-informed curriculum

The human side, built and ready

This is the scalability argument the white paper has been building toward. The problem isn't that peer support networks are hard to create. It's that most organizations attempting to integrate AI with human support don't know how to build the human side. STAR Network does.

Within 60 days of a licensing agreement, STAR Network can train mentors, deploy a virtual meeting schedule, and begin collecting outcome data — in any language or geography, and at a per-person cost that is negligible compared to either professional therapy or the downstream costs of untreated trauma.



First applications beyond TAR Anon



The model is already in early deployment with law enforcement agencies and is being piloted in partnership with AI platforms including *My What If* and *Wingman*.

These are the first applications beyond TAR Anon, and will not be the last.

References

1. American Psychological Association. (2023). 2023 COVID-19 Practitioner Survey. APA.
2. American Psychological Association. (2025, November). Health Advisory on Generative AI Chatbots and Wellness Applications for Mental Health. APA.
3. American Psychological Association. (2025, November). Stress in America 2025: A Crisis of Connection. APA.
4. Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, 16(3), 252–260.
5. Bureau of Health Workforce, HRSA. (2023). Behavioral Health Workforce Brief.
6. De Freitas, J., Oğuz-Uğuralp, Z., & Kaan-Uğuralp, A. (2025). Emotional Manipulation by AI Companions. HBS Working Paper No. 26-005.
7. Dupre, M. H. (2025, June 28). People are being involuntarily committed after spiraling into “ChatGPT psychosis.” *Futurism*.
8. Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886.
9. Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction. *Psychological Science*, 19(2), 114–120.
10. Fang, M. et al. (2025). How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use. MIT Media Lab / OpenAI.
11. Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy. *Psychotherapy*, 55(4), 316–340.
12. George Mason University College of Public Health. (2025, December). New survey on AI for managing stress, anxiety, and mental health needs.
13. Haber, N. et al. (2025). Exploring the Dangers of AI in Mental Health Care. Stanford HAI / ACM FAccT.
14. Hao, K. et al. (2025). AI companion response adequacy study. Stanford / CMU / UMN.
15. McBain, R. K. et al. (2025). Use of Generative AI for Mental Health Advice Among US Adolescents and Young Adults. *JAMA Network Open*, 8(11), e2542281.
16. National Council for Mental Wellbeing. (2023). Behavioral health workforce shortage study.
17. OpenAI. (2025, October). Mental health and safety disclosures.
18. Østergaard, S. D. (2023). Chatbot-triggered psychosis. *Schizophrenia Bulletin*, 49(6).
19. Payne, K. (2024, October 25). An AI chatbot pushed a teen to kill himself. *AP News*.
20. Prinstein, M. (2025). Examining the Harm of AI Chatbots. Testimony before the U.S. Senate Judiciary Committee. APA.
21. Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources. *Practice Innovations*. DOI: 10.1037/pri0000292.
22. Sakata, K. (2025). AI psychosis: Clinical observations. UCSF.
23. Turkle, S. (2011). *Alone Together*. Basic Books.
24. Weizenbaum, J. (1976). *Computer Power and Human Reason*. W. H. Freeman.
25. Yao, M. et al. (2025). Connecting Self-Esteem to Problematic AI Chatbot Use. *Frontiers in Psychology*.
26. Yin, Y., Jia, N., & Waksak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *PNAS*, 121(14), e2319112121.

PARTNER WITH STAR NETWORK

Our peer-support system is ready to deploy.

The organizations that integrate it first will set the standard.

Reach out, and let's set the standard together.

CONTACT

contact@starnetwork.org · starnetwork.org

Trauma ends where connection begins.