Ashleigh Golden, PsyD • September 2025

wayhaven

Safety and ethics in AI mental wellness

Building on our commitments

Safety and ethics have always been central

Over the past several months, headlines about “AI psychosis,” youth suicidality linked to LLM-based chatbots, and emerging lawsuits have raised understandable alarm (Roose, 2024; Klee, 2025; Hart, 2025; Hill, 2025; Rosenberg, 2025). The throughline is not that AI itself is inherently harmful, but that tools not built for mental wellness and certainly not tested for it are being used in ways that outstrip their design. A tool being fit for purpose matters. When systems lack boundaries, guardrails, and accountability, they can foster dependency, blur reality, and fail people at the moment that they most need human connection. As Steve Duke (2025) observes, there is now a “moral imperative” to build safer alternatives designed by clinical experts, grounded in ethics, user-centered in design, equipped with safety protocols, and demonstrably delivering outcomes.

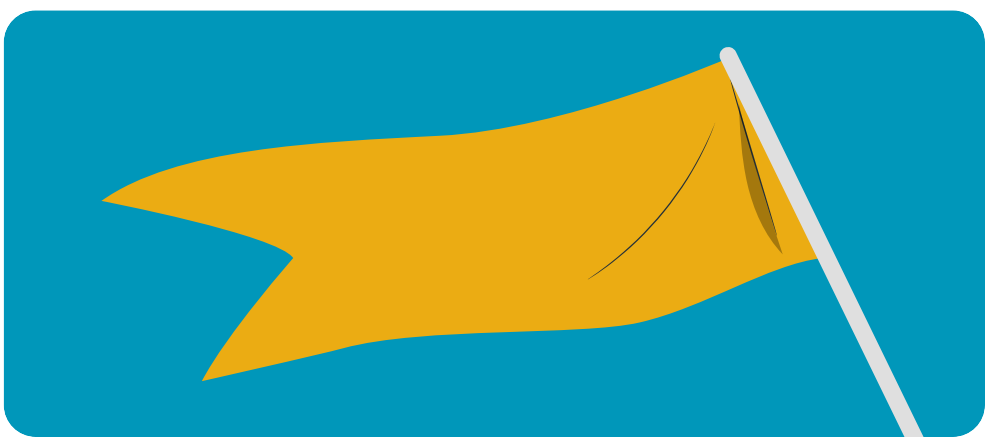
“**A tool being fit for purpose matters.** When systems lack boundaries, guardrails, and accountability, they can foster dependency, blur reality, and fail people at the moment that they **most need human connection.**”

Wayhaven was founded on exactly this premise: a fit-for-purpose AI mental wellness tool with bounded scope, transparent role, and safeguards that are native to the product, not bolted on. Indeed, more than a year ago, we published our ethics commitments to informed consent, clear disclosure that users are engaging with AI coaches, capability and limitation framing, privacy controls, evidence-informed mechanisms, crisis protocols, and user empowerment (Golden, 2024). I also recently spoke to The Washington Post in its explainer, which underscored that guidance is needed now, not later (Tiku & Malhi, 2025). In addition, the policy environment is shifting rapidly. In August 2025, Illinois became the first state to ban the use of AI for independent therapeutic decision-making without licensed clinician oversight, with Governor Pritzker signing the Wellness and Oversight for Psychological Resources Act into law (Blum, 2025).

These concerns are exactly why structured evaluation frameworks matter. The Framework for AI Tool Assessment in Mental Health (FAITA-MH) was developed as a public-facing scale to help clinicians, developers, and the public assess AI mental wellness products across domains such as credibility, user experience, crisis protocols, user agency, and transparency (Golden & Aboujaoude, 2024).

The Readiness Evaluation for AI-Mental Health Deployment and Implementation (READI) framework was designed to guide responsible deployment by attending to similar criteria while adding an implementation dimension (Stade, Eichstaedt, Kim, & Wiltsey Stirman, 2025). Taken together, FAITA-MH provides a structured lens for product-level safeguards and disclosures, while READI adds an implementation-oriented lens for how AI mental wellness technologies are deployed and sustained in practice.

This makes it timely to reaffirm how Wayhaven differentiates itself among fit-for-purpose AI mental wellness apps, in terms of safety, ethics, effectiveness, feasibility, and acceptability, as well as in how we design our AI coaches for meaningful engagement and responsible use.



Building on our foundation: Progress and priorities

What the evidence shows

Stade, Stirman, et al. (2024) emphasize that assessing feasibility and acceptability is a necessary stage in digital mental health evaluation, and the Readiness Evaluation for AI-Mental Health Deployment and Implementation (READI) framework highlights the importance of examining these factors for all stakeholders (Stade, Eichstaedt, Kim & Stirman, 2024). User experience, feasibility, and acceptability must be assessed before moving on to larger outcome trials (Stade, Stirman, et al., 2024).

In July 2025, we published Generative AI-Powered Mental Wellness Chatbot for College Student Mental Wellness: Open Trial in JMIR Formative Research in collaboration with the Reyes-Portillo Lab at Montclair State University and the Fit Minded team, headed by Jennifer Huberty, former Head of Science at Calm (Reyes-Portillo et al., 2025). The study focused on end users as the primary stakeholder group and explored preliminary efficacy indicators. Participants were 50 Montclair State college students, a majority of whom self-identified as racially and ethnically diverse, with elevated anxiety or depression symptoms.

Eligible participants completed pre- and post-surveys after spending at least 5 minutes engaged in a Wayhaven conversation, then were given the option to use Wayhaven over the course of a week with encouragement to engage as much as needed.

Early outcomes

Results showed statistically significant decreases in feelings of depression, anxiety, and hopelessness, alongside significant increases in agency, self-efficacy, and well-being across the study period.

Feasibility and acceptability

At follow-up, 90% of participants found Wayhaven easy to use and 80% said it fit easily into their daily routine. Most students expressed satisfaction, agreed that they would use it again, endorsed the need for a resource like Wayhaven among university students, and said that they would recommend it to a peer.

How students used Wayhaven

In terms of actual use, beyond the required initial 5-minute session, participants had optional access to Wayhaven for one week. During this period, they completed an average of two sessions, exchanged about 29 messages per session, and spent roughly 15 minutes in each session with their chosen AI coach. Engagement is important insofar as it provides sufficient exposure to active ingredients and mechanisms of change to contribute to meaningful outcomes (Stade, Stirman, et al., 2024). On average, students also practiced about one CBT-based skill per session, such as grounding, mindfulness, or cognitive restructuring, which indicates that engagement extended beyond conversation into interaction with core active ingredients.

Building on our foundation: Progress and priorities (continued)

Taken together, these findings suggest that Wayhaven is a feasible, acceptable, and engaging resource for diverse college students, including those with clinically elevated depression and anxiety symptoms. Reyes-Portillo et al. (2025) conclude that Wayhaven may represent a pathway to accessible, free, equitable mental wellness support for a population with overwhelming unmet needs. These early results indicate that Wayhaven can sustain engagement for long enough to deliver core skills and to begin shifting outcomes.

As Duke (2025) stresses, the only way to shift behavior away from generic conversational AIs already being used for well-being is to demonstrate that purpose-built tools actually improve people's well-being and to develop robust, compelling evidence of that impact. This imperative underscores both the significance of early efficacy indicators from the open trial and the need for more rigorous outcomes research.

Next steps

As Stade, Stirman, et al. (2024) recommend, the next stage in evaluation will be rigorous outcomes testing. The Reyes-Portillo lab is analyzing the qualitative data set from the open trial (Reyes-Portillo et al., 2025), with plans to publish later this year. In parallel, the Wayhaven–Reyes-Portillo–Fit Minded team is preparing for the next phase of evaluation, which will use more robust study designs to extend findings beyond the open trial.

Real-world engagement: How students are using Wayhaven in the wild

Findings from the JMIR open trial (Reyes-Portillo et al., 2025) showed that students found Wayhaven both feasible and acceptable, and real-world use since launch suggests that the same holds outside of a study context. For example, after chats, students can provide a 5-star rating of helpfulness, serving as a proxy indicator of acceptability. Since launch, 80% of those ratings have been “helpful” or “very helpful.”

Among Wayhaven’s native app users, 70% return for more than one conversation and 51% have more than four conversations. Across all users, 30% complete four or more chats, with an average of 2.8 chats per user. Within those conversations, students exchange about 12 messages on average, for a total of roughly 34 messages per user.

80%

Percent of chats rated as "helpful" or "very helpful

70%

Percent of app users who have multiple chats

34

Number of messages per chat user

These naturalistic data suggest that students not only try Wayhaven but also return, often multiple times, and engage deeply enough to sustain dozens of conversational turns. However, we should remind ourselves that engagement is important insofar as it provides exposure to key active ingredients from behavioral science, but engagement on its own is not sufficient to produce meaningful change (Stade, Stirman, et al., 2024). What remains unknown is the precise relationship between engagement and outcomes; i.e., do the largest gains occur in the first few conversations, do benefits continue to mount with additional use, or does later use primarily reinforce early gains?

Even with those uncertainties, early efficacy indicators from the JMIR open trial (Reyes-Portillo et al., 2025) showed improvements in mood, hopelessness, agency, and well-being after even brief engagement, suggesting that meaningful change may occur early. Real-world patterns, meanwhile, indicate that students are willing to return and participate in skill learning and practice repeatedly, providing additional opportunities for reinforcement. Clarifying this “dose-response” relationship is an area for future research and will be important to explore in more rigorous studies. Rather than measuring success purely by maximizing minutes or endless engagement, as some generic AI models might, our design is grounded in the principle that engagement is valuable only when it contributes to well-being, not as an endpoint in itself.

Duke (2025) also notes that the central challenge in AI for mental well-being is sustaining engagement without compromising evidence-informed design principles. Even the safest tools have little impact if students do not use them, which is why engagement must go hand in hand with responsible design.

These findings also align with multiple engagement indicators highlighted in the READI framework, including application use metrics such as average time per conversation and number of chats, as well as user ratings that can serve as a proxy for satisfaction (Stade, Eichstaedt, Kim, & Stirman, 2025). Other indicators, such as alliance measures, remain areas for future research.

Safety and crisis handling

Wayhaven's approach to safety is anchored in FAITA-MH's safety and crisis management dimension and READI's safety criterion (Golden & Aboujaoude, 2024; Stade, Eichstaedt, Kim, & Stirman, 2025). We detect and route for situations involving potential harm to self, harm to others, harm from others, and medical emergencies. Human support is always prioritized, because safety in possible crisis situations ultimately depends on connecting students to people and systems trained and equipped to provide tailored care. Conversations are governed by clear rules about when they must end, making sure that students are encouraged toward real-world support rather than remaining in prolonged chats that could become avoidant. Within those limits, we integrate evidence-informed distress tolerance skills that provide short-term relief and help students take the next step toward human resources. Action planning reinforces this process by guiding students to identify one concrete step and how they will carry it out, while also surfacing and addressing common barriers to follow-through with human support.

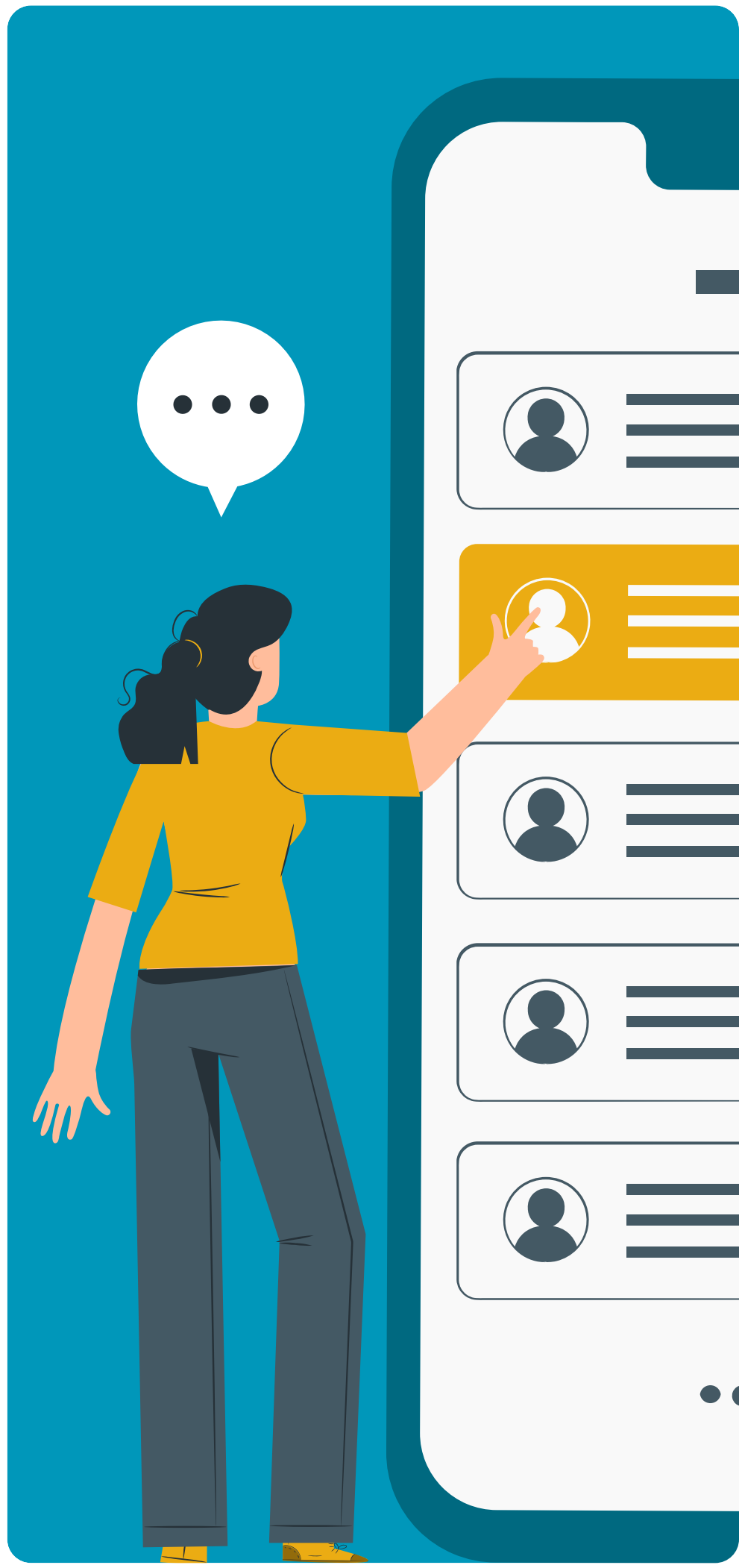
Guardrails guide the AI to validate and affirm appropriately without sycophancy, a tendency to over-agree or flatter users in ways that may reinforce maladaptive beliefs and behaviors (Dohnány et al., 2025). This helps conversations focus on strengthening skills-building and social connection rather than reinforce unhealthy patterns.

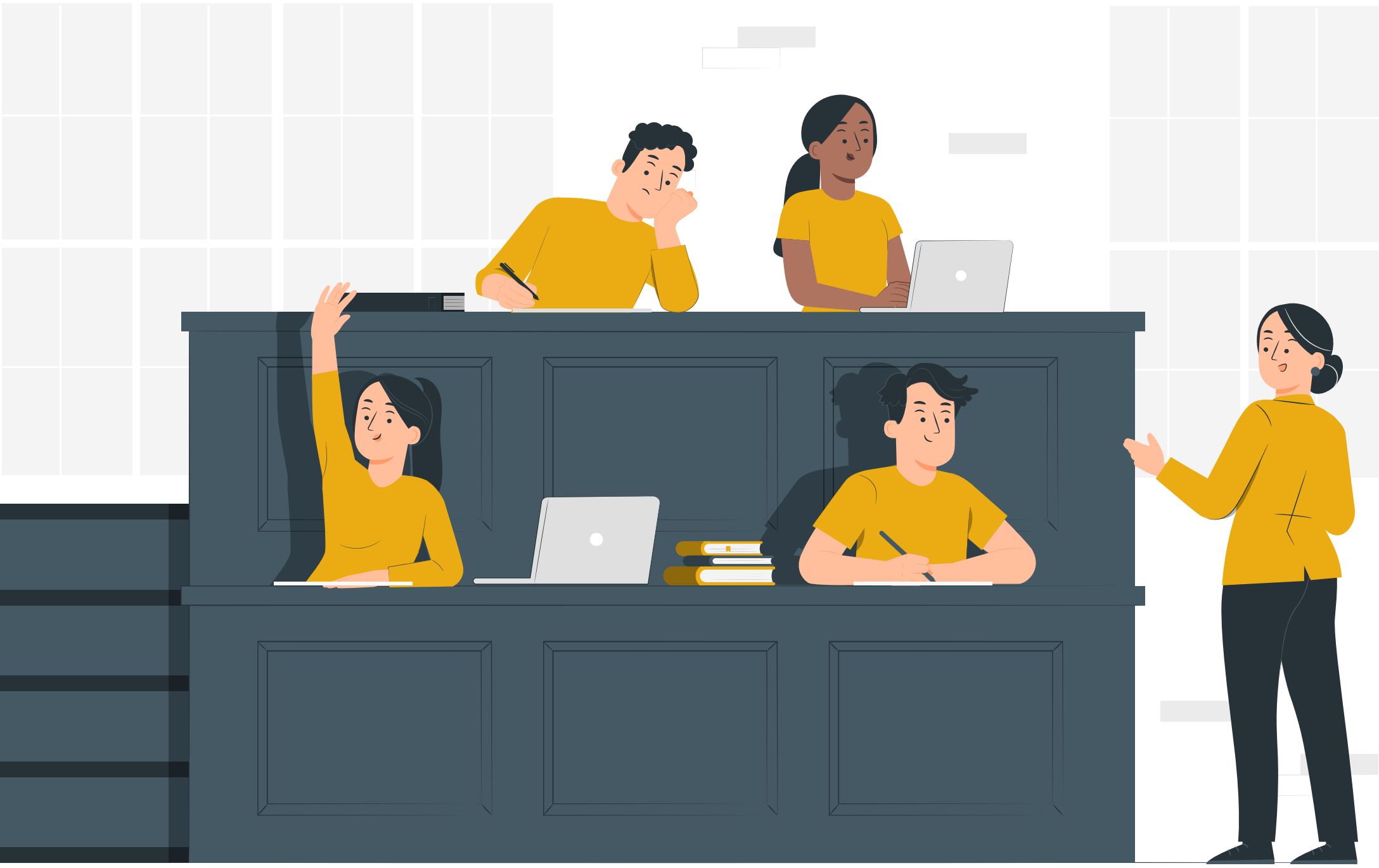
"**Human support is always prioritized,** because safety in possible crisis situations ultimately depends on connecting students to people and systems trained and equipped to provide tailored care."

AI coach design: Engagement, guardrails, and responsible use

Beyond crisis protocols, AI coach design itself is a key dimension of safety and ethics. Evidence suggests that personality is an important factor in the effectiveness of mental health chatbots (Moilanen et al., 2022), and that anthropomorphism, the tendency to attribute agency, intentionality, emotional states, or even consciousness to systems exhibiting complex behavior (Dohnány et al., 2025), is one of the most important drivers of acceptability and adoption of conversational agents in health care (Wutz et al., 2023). Consistent with best practices in conversational design for mental well-being, some degree of personification may therefore play a meaningful role in design. Our UXR demonstrates that students prefer AI coaches who reflect their own diverse identities and lived experiences, finding such AI coaches more relatable, trustworthy, and motivating. At the same time, safeguards are key to prevent anthropomorphism from contributing to cycles that could reinforce maladaptive beliefs or behaviors. That makes it essential to set clear boundaries that preserve safety, transparency, and appropriate use of the tool.

Recently, concerns have been raised about anthropomorphism in the context of generic large language models, particularly when people use them for mental well-being support. Our approach is grounded in a careful, ethical balance of two lines of thinking: first, best practices in conversation design for mental well-being and user research that support the value of relatable, identity-reflective AI coaches, and second, emerging mechanistic models and safety frameworks highlight the need for guardrails to prevent dependence and to counter mechanisms such as sycophancy that can reinforce maladaptive beliefs and behaviors (Morrin et al., 2025; Dohnány et al., 2025). AI coaches should be engaging enough to foster trust and uptake to increase the probability that students interact with key active ingredients and achieve meaningful outcomes, but not designed in ways that blur the line between simulation and reality. As cyberpsychologist Rachel Wood has noted, personas can add value when they are intentionally designed and paired with clear reminders of non-personhood, but risk creating a false sense of intimacy if those boundaries are not reinforced (Wood, 2025).





Our ongoing commitment

At Wayhaven, user safety and responsible innovation are the foundation of how we support students and partner with campus leaders, while offering parents and caregivers reassurance that their communities have access to a safe alternative. In a time when many AI tools are drawing concern for being unbounded or untested, our focus remains on offering AI mental wellness coaches that people can use with confidence. We ground our work in frameworks such as FAITA-MH (Golden & Aboujaoude, 2024) and READI (Stade, Eichstaedt, Kim, & Stirman, 2025), and we stay current with best practice guidelines and the academic and clinical literature so that our design decisions are tied to evidence, not intuition. That balance includes making sure that our AI coaches remain engaging enough to foster trust and skill practice, while bounded by safeguards that prevent dependence or blurred reality, or reinforcement of maladaptive patterns.

By keeping well-being front of mind and refining our safeguards as new research and standards emerge, we aim to sustain and strengthen Wayhaven as a safe and trusted space for building skills and finding meaningful support. Transparency and fit-for-purpose design will continue to guide our work as both the technology and the needs of the communities we serve evolve.

References

Blum, K. (2025, August 20). States crack down on AI for behavioral health care. Health Equity. https://healthequity.com/article/states-crack-down-ai-behavioral-health

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Summerfield, C., ... & Nour, M. M. (2025). Technological folie à deux: Feedback loops between AI chatbots and mental illness. arXiv preprint. https://doi.org/10.48550/arXiv.2507.19218

Duke, S. (2025, September 5). Some thoughts on conversational AI: The moral imperative for mental health organisations to build better, safer AI agents, and why time is running out [Newsletter]. The Hemingway Report. https://thehemingwayreport.beehiiv.com

Golden, A. (2024, August 13). Ethical AI: Upholding best practices in responsible mental wellness AI development. Wayhaven. https://www.wayhaven.com/post/ethical-ai

Golden, A., & Aboujaoude, E. (2024). Describing the Framework for AI Tool Assessment in Mental Health and applying it to a generative AI obsessive-compulsive disorder platform: Tutorial. JMIR Formative Research, 8(1), e62963. https://doi.org/10.2196/62963

Golden, A., & Aboujaoude, E. (2024). The Framework for AI Tool Assessment in Mental Health (FAITA-Mental Health): A scale for evaluating AI-powered mental health tools. World Psychiatry, 23(3), 444–445. https://doi.org/10.1002/wps.21248

Hart, R. (2025, August 5). Chatbots can trigger a mental health crisis. What to know about “AI psychosis.” TIME. https://time.com/7307589/ai-psychosis-chatgpt-mental-health

Hill, K. (2025, June 13) They asked A.I. chatbots questions. The answers sent them spiraling. New York Times. https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html

Illinois Department of Financial and Professional Regulation. (2025, August 4). Gov. Pritzker signs legislation prohibiting AI therapy in Illinois [Press release]. https://idfpr.illinois.gov/content/dam/soi/en/web/idfpr/news/2025/2025-08-04-idfpr-press-release-hb1806.pdf?utm_source=chatgpt.com

Klee, M. (2025, May 4). People are losing loved ones to AI-fueled spiritual fantasies. Rolling Stone. https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/

Moilanen, J., Visuri, A., Suryanarayana, S. A., Alorwu, A., Yatani, K., & Hosio, S. (2022, November). Measuring the effect of mental health chatbot personality on user engagement. In Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia (pp. 138-150). https://doi.org/10.1145/3568444.3568464

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ... & Pollak, T. A. (2025). Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). PsyArXiv. https://doi.org/10.31234/osf.io/cmy7n_v5

References (continued)

Rosenberg, S. (2025, August 26). Parents sue OpenAI over teen's suicide. Axios. https://www.axios.com/2025/08/26/parents-sue-openai-chatgpt

Reyes-Portillo, J. A., So, A., McAlister, K., Nicodemus, C., Golden, A., Jacobson, C., & Huberty, J. (2025). Generative AI–Powered Mental Wellness Chatbot for College Student Mental Wellness: Open Trial. JMIR Formative Research, 9(1), e71923. https://doi.org/10.2196/71923

Roose, K. (2024, October 23). Can A.I. be blamed for a teen's suicide? The New York Times. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html

Stade, B., Eichstaedt, J. C., Kim, J. P., & Stirman, S. W. (2025). Readiness evaluation for AI–mental health deployment and implementation (READI): A review and proposed framework. Technology, Mind, and Behavior. https://doi.org/10.1037/tmb0000163

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., ... & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. NPJ Mental Health Research, 3(1), 12. https://doi.org/10.1038/s44184-024-00056-z

Tiku, N., & Malhi, S. (2025, August 19). What is "AI psychosis" and how can ChatGPT affect your mental health? The Washington Post. https://www.washingtonpost.com/health/2025/08/19/ai-psychosis-chatgpt-explained-mental-health/

Wood, R. (2025). Comment on AI personas and non-personhood [LinkedIn post]. LinkedIn. https://www.linkedin.com/feed/update/urn:li:activity:7366453829250289664?commentUrn=urn%3Ali%3Acomment%3A%28activity%3A7366453829250289664%2C7366624503356903426%29&replyUrn=urn%3Ali%3Acomment%3A%28activity%3A7366453829250289664%2C7366643251623178241%29&dashCommentUrn=urn%3Ali%3Afsd_comment%3A%287366624503356903426%2Curn%3Ali%3Aactivity%3A7366453829250289664%29&dashReplyUrn=urn%3Ali%3Afsd_comment%3A%287366643251623178241%2Curn%3Ali%3Aactivity%3A7366453829250289664%29

Wutz, M., Hermes, M., Winter, V., & Köberlein-Neu, J. (2023). Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: Integrative review. Journal of Medical Internet Research, 25, e46548. https://doi.org/10.2196/46548