IJBMC International Journal of Body, Mind and Culture

Volume 13, Issue 11, Article e2026-1779
DOI: 10.61838/ijbmc.v13i11.1779

Article type:
Editorial

Relational Safety in Generative AI Mental Health Care

1 Department of Psychology and Counseling, KMAN Research Institute, Richmond Hill, Ontario, Canada.

2 PhD in Clinical Psychology, Lecturer, Islamic Azad University, Tehran Medical Sciences Branch, Tehran, Iran.

Corresponding author:
hadialitabar@kmanresce.ca

Seyed Hadi. Seyed Alitabar[1*], Hasti. Salehi[2]



Article history:

Received 22 Sep 2026
Revised 18 Oct 2026
Accepted 28 Oct 2026
Published online 01 Nov 2026

How to cite this article:
Seyed Alitabar, S. H., & Salehi, H. (2026). Relational Safety in Generative AI Mental Health Care. International Journal of Body, Mind and Culture, 13(11), Article e2026-1779. https://doi.org/10.61838/ijbmc.v13i11.1779

© 2026 the authors. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

ABSTRACT

Generative artificial intelligence may provide psychological support, but favourable ratings do not establish safety. We define relational safety as preserving autonomy, avoiding reinforcement of false and harmful beliefs, preventing dependence that impairs functioning, and facilitating appropriate human support over time. This interactional construct complements clinical harm prevention and differs from perceived therapeutic alliance. Trials suggest benefits for specific interventions; experiments and simulations identify potential risks without establishing their population incidence. We propose four complementary domains addressing judgement, dependence and social connection, continuity of care, and embodied cultural meaning. Candidate indicators include belief endorsement, impaired control, completed referrals, and misunderstanding of bodily distress. Symptoms, functioning, and autonomy should be reported separately. Accountable escalation and transparent data and memory practices are necessary evaluation targets. These proposals require prospective validation; neither frequent use nor satisfaction alone demonstrates benefit or harmful dependence.

Keywords: Generative artificial intelligence, Mental health, Relational safety, Therapeutic alliance, Cultural adaptation, Patient autonomy.

Why relational safety matters

Generative AI technologies should be evaluated for their effects on decision making, relationships, and access to support. We define relational safety as an interaction's capacity, over time, to preserve autonomy, avoid reinforcing false and harmful (dysfunctional) beliefs, prevent dependence that impairs functioning, and facilitate appropriate human support. Clinical safety concerns prevention and management of health-related harm; therapeutic alliance concerns the perceived bond and agreement on goals and tasks. Relational safety overlaps with these constructs but specifically examines evolving patterns of reliance, belief reinforcement, and connection to care.

Therapeutic potential needs consideration. In a randomised trial consisting of 210 adults, symptom improvement was greater in participants who used the specialised generative chatbot compared to those on waitlist at 4 and 8 weeks (Heinz et al., 2025). However, this demonstrates therapeutic potential under these specific circumstances of this intervention rather than equating it to psychotherapy or generalising to all consumer chatbots. Although there is an association between alliance and psychotherapy outcomes (Flückiger et al., 2018), positive ratings of alliance would not indicate clinical skill. The importance of human connection and embodiment should still be considered (Seyed Alitabar, 2026).

When affirmation becomes a risk

Pre-registered studies of 2,405 participants found that sycophantic AI reduced responsibility-taking and intentions to repair relationships despite being preferred and trusted (Cheng et al., 2026). These were judgements and intentions, not psychiatric incidence. Training models for warmth may also reduce accuracy and increase affirmation of incorrect beliefs (Ibrahim et al., 2026). Empathy acknowledges feelings without endorsing an unverified explanation or harmful belief. This boundary applies to human relationships too: a response can validate distress while expressing uncertainty and respectfully questioning its interpretation.

Vulnerability-amplifying feedback loops may arise through interactions between user vulnerabilities and chatbot behaviour (Dohnány et al., 2026). Across 810 simulated conversations, SIM-VAIL identified concerning behaviour that varied by user profile and intent and could accumulate (Weilnhammer et al., 2026). Simulation identifies failure modes, not real-world incidence. A scoping review also distinguished hypothetical and observed harms while acknowledging causal uncertainty (Diel et al., 2026). Prospective studies need exposure denominators, baseline vulnerability assessments, and independent adverse-event review.

Evaluating change outside the conversation

Table 1 operationalises relational safety alongside existing lifecycle accountability frameworks (Lee et al., 2026). Its contribution is an evaluative synthesis, not a claim that the component risks are new. The four domains address distinct pathways: distorted judgement, displacement of human relationships, failed transitions to care, and misinterpretation of bodily or cultural experience. They are complementary and potentially overlapping, rather than exhaustive or validated dimensions.

Table 1

Proposed domains and candidate outcomes for relational safety

Domain	Proposed indicators	Assessment
Epistemic agency	False-belief endorsement rate; change in autonomous decision making	Blinded ratings of standardised multi-turn scenarios; locally validated autonomy measures
Dependence and social connection	Impaired control; interrupted-access distress; displaced activities or human contact	Baseline and follow-up measures of impairment alongside use frequency and goal attainment

Domain	Proposed indicators	Assessment
Continuity of care	Completed referrals / indicated referrals; time to qualified assessment; missed or unnecessary escalation	Audit eligible events and handover completion against prespecified response windows
Embodied cultural meaning	Misinterpretation rate for bodily idioms; sleep disruption; bodily distress and activity interference	Local scenario review; repeated patient reports; language-specific equivalence and subgroup checks

Note. These are candidate indicators, not validated thresholds. Define events, denominators, time windows, and clinically meaningful change prospectively. Emotional validation must be distinguished from belief endorsement; cultural difference alone does not establish dysfunction.

Culture shapes reassurance, bodily distress, privacy, and help-seeking (Seyed Alitabar, 2025). Cultural adaptation evidence supports local evaluation, not automatic validation of AI adaptations (Hall et al., 2016). Cultural formulation contextualises symptoms and personal explanations (Kirmayer & Ryder, 2016). Local clinicians and people with lived experience should assess idioms of distress, spiritual meanings, and acceptable disagreement, avoiding cultural stereotypes or misclassification of unfamiliar beliefs as psychopathology.

Embodiment concerns how bodily sensations contribute to emotional experience, beyond verbal symptom descriptions (Seth, 2013). Candidate indicators include sleep disturbance, bodily distress, repetitive symptom checking, and interference with daily activity. Language-specific tests should assess misunderstanding of bodily idioms and distinguish emotional support from inappropriate reassurance about physical symptoms. Assess measurement equivalence and subgroup uncertainty; cultural responsiveness must preserve personal choice and safety.

A testable research and governance agenda

Trials should specify interventions and credible active comparators with similar attention, accessibility, and therapeutic content. Frequent, purposeful use may remain beneficial when control, functioning, and human support are preserved. Harmful dependence involves impaired control, distress around interrupted access, or displacement of valued activity and support; frequency alone is insufficient. Validated measures and follow-up after access ends should distinguish these patterns.

Record model versions, instructions, memory settings, and product changes. DECIDE-AI addresses contextual and human factors in decision support, with limited direct applicability to autonomous therapeutic chatbots (Vasey et al., 2022). Independently audit conversation trajectories and disagreements with automated safety scores. Test safeguards and support for offline goals rather than assuming effectiveness.

Confidentiality requires data minimisation, access controls, and clear disclosure of retention, deletion, and secondary uses. Persistent memory should be optional, reviewable, and deletable, with its limits explained. Specify who can access conversations and whether data inform model training; obtain explicit consent for secondary use rather than inferring it from service use. Safety audits should use de-identified data where possible.

Health services should define intended use, evidence requirements, and escalation responsibility, consistent with WHO governance guidance (World Health Organization, 2024). Developers implement detection and escalation pathways; deploying services designate qualified clinicians, response times, and fallback arrangements. Worsening symptoms warrant human assessment; imminent danger requires a verified local emergency pathway. Disclose when monitoring is unavailable. With appropriate consent and confidentiality safeguards, document referrals, time to contact, completed handovers, false reassurance, and unnecessary escalation.

Report symptoms, functioning, autonomy, and adverse events separately, so improvement in one cannot conceal deterioration in another. Prespecify clinically meaningful change with patients and clinicians; report missing outcomes and follow up discontinuers. Any joint success criterion should retain these separate results and requires prospective testing.

Ijbmc.org
E-ISSN: 2345-5802

Conclusion

Relational safety may complement clinical safety and alliance assessment by examining autonomy, reliance, and access to human care over time. The proposed domains and indicators require prospective validation across cultural and clinical contexts. Current evidence supports investigation, not a conclusion that these safeguards establish safety or effectiveness.

References

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, *391*(6792), Article eaec8352. https://doi.org/10.1126/science.aec8352

Diel, A., Torous, J., Cuijpers, P., Kleesiek, J., Nensa, F., Weber, N., Faust, F., Lalgi, T. J., Mellis, F. S., Teufel, M., & Bäuerle, A. (2026). A scoping review on the mental health harms of LLM-based chatbots. *npj Digital Medicine*, *9*(1), Article 644. https://doi.org/10.1038/s41746-026-03054-x

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., & Nour, M. M. (2026). Technological folie à deux: Feedback loops between AI chatbots and mental health. *Nature Mental Health*, *4*(3), 336–345. https://doi.org/10.1038/s44220-026-00595-8

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, *55*(4), 316–340. https://doi.org/10.1037/pst0000172

Hall, G. C. N., Ibaraki, A. Y., Huang, E. R., Marti, C. N., & Stice, E. (2016). A meta-analysis of cultural adaptations of psychological interventions. *Behavior Therapy*, *47*(6), 993–1014. https://doi.org/10.1016/j.beth.2016.09.005

Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*, *2*(4), Article AIoa2400802. https://doi.org/10.1056/AIoa2400802

Ibrahim, L., Hafner, F. S., & Rocher, L. (2026). Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, *652*(8112), 1159–1165. https://doi.org/10.1038/s41586-026-10410-0

Kirmayer, L. J., & Ryder, A. G. (2016). Culture and psychopathology. *Current Opinion in Psychology*, *8*, 143–148. https://doi.org/10.1016/j.copsyc.2015.10.020

Lee, H., Handler, R., Mungle, T., & Hernandez-Boussard, T. (2026). Building safer artificial intelligence mental health chatbots: A framework for transparency, evaluation, and shared accountability. *Journal of the American Medical Informatics Association*, *33*(8), 1538–1553. https://doi.org/10.1093/jamia/ocag078

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Seyed Alitabar, S. H. (2025). The cultural imprint on the mind–body connection: Rethinking health beyond the biomedical model. *International Journal of Body, Mind and Culture*, *12*(4), 1–4. https://doi.org/10.61838/ijbmc.v12i4.1086

Seyed Alitabar, S. H. (2026). The forgotten body in digital mental health: Why culture, embodiment, and human connection must shape the future of AI-assisted care. *International Journal of Body, Mind and Culture*, *13*(6), Article e2026-1624. https://doi.org/10.61838/ijbmc.v13i6.1624

Vasey, B., Nagendran, M., Campbell, B., Clifton, D. A., Collins, G. S., Denaxas, S., Denniston, A. K., Faes, L., Geerts, B., Ibrahim, M., Liu, X., Mateen, B. A., Mathur, P., McCradden, M. D., Morgan, L., Ordish, J., Rogers, C., Saria, S., Ting, D. S. W., … DECIDE-AI expert group. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*, *28*(5), 924–933. https://doi.org/10.1038/s41591-022-01772-9

Weilnhammer, V., Hou, K. Y. C., Luettgau, L., Summerfield, C., Dolan, R., & Nour, M. M. (2026). A clinically validated framework for auditing AI chatbot behavior in mental health interactions. *Nature Medicine*, *32*(10), 3730–3740. https://doi.org/10.1038/s41591-026-04577-2

World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. https://www.who.int/publications/i/item/9789240084759