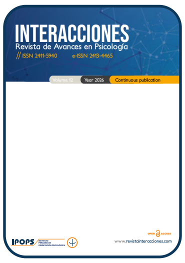
INTERACCIONES
Journal of family, clinical and health psychology
// ISSN 2411-5940 e-ISSN 2413-4465

www.revistainteracciones.com



EDITORIAL

Ethical and Regulatory Gaps in Using Generative AI for Mental Health Support in Low- and Middle-Income Countries

Leonardo Rojas-Mezarina[1 *], David Villarreal-Zegarra[2]
[1] Facultad de Medicina, Universidad Nacional Mayor de San Marcos, Lima, Peru.
[2] Digital Health Research Center, Instituto Peruano de Orientación Psicológica, Lima, Peru.

*** Correspondence:** leonardo.rojas@unmsm.edu.pe.

Received: December 13, 2025 | **Reviewed:** December 27, 2025 | **Accepted:** January 02, 2026 | **Published Online:** January 02, 2026.

CITE IT AS:
Rojas-Mezarina, L., Villarreal-Zegarra, D. (2026). Ethical and Regulatory Gaps in Using Generative AI for Mental Health Support in Low- and Middle-Income Countries. *Interacciones, 12*, e497. http://dx.doi.org/10.24016/2026.v12.497

The accelerated adoption of generative artificial intelligence (AI) models, such as ChatGPT and Gemini, as well as other conversational agents, has transformed how people worldwide seek mental health information and support (Thirunavukarasu et al., 2023). These large language models (LLMs) are being used at massive scale; for example, ChatGPT alone has been reported to have hundreds of millions of weekly active users (Chatterji et al., 2025). Users interact with these systems to receive guidance related to anxiety, depression, or crisis situations, marking an unprecedented shift in the digital health ecosystem (Ayers et al., 2023). However, while generative AI promises to expand access to physical and mental health resources, it also introduces ethical and regulatory risks that remain insufficiently addressed (Meskó & Topol, 2023), particularly in low- and middle-income regions such as Latin America. In these settings, AI models developed in high-income countries are widely deployed without necessarily assessing the potential risks of bias that this entails (Hussain et al., 2025).

In low- and middle-income countries (LMICs), the governance architecture for health-related generative artificial intelligence, encompassing standards, accountability, transparency, and enforceable data protection, lags behind its real-world implementation. We observe that AI-based systems are increasingly integrated into daily life without adequate standards for safety, transparency, or data protection (Morley et al., 2020). The risks arising from their therapeutic or quasi-therapeutic use in mental health therefore warrant urgent examination.

First, using LLMs for mental health support entails processing intimate and highly sensitive information, including symptoms, trauma narratives, medication histories, and crisis-related disclosures (Mandal et al., 2025; Wang et al., 2025). These interactions can also generate sensitive inferences (e.g., suicide risk, substance use, or exposure to abuse) even when users do not explicitly disclose them, increasing the potential for privacy harms if data are mishandled.

Second, the technological infrastructure that enables these services is commonly located outside the jurisdictions of LMICs, under privacy policies that permit the use, storage, and training on personal user data (Vollmer et al., 2020). This extraterritoriality complicates enforcement and redress mechanisms and weakens cross-border accountability, particularly where local regulatory agencies have limited technical capacity or unclear legal authority over foreign providers.

Third, foundational model development and data management remain opaque, including uncertainty regarding the provenance of training corpora, data governance practices, and safeguards to meet expectations of medical confidentiality (Bommasani et al., 2023). The "black box" nature of these systems also complicates auditability and post hoc investigation when harmful outputs occur, limiting effective oversight (Ethical AI governance group, 2023).

In the Peruvian case, the Law on Personal Data Protection (Law No. 29733) is insufficient to address emerging generative AI risks because it does not encompass critical aspects such

Publication edited by the Instituto Peruano de Orientación Psicológica - IPOPS

https://doi.org/10.24016/2026.v12.497

as sensitive inferences, algorithmic reuse, or re-identification risks (Smart & Montori, 2025). This gap is particularly salient because, while Peru's data protection framework shares broad intent with comprehensive regimes such as the EU's GDPR, it is not directly comparable to sector-specific U.S. frameworks such as HIPAA and does not yet address AI-specific risks (e.g., sensitive inferences, algorithmic reuse, and re-identification). As a result, users may be exposed to privacy violations with emotional, clinical, and societal consequences.

Generative AI models are predominantly trained on data in English and within Western sociocultural contexts, and may generate erroneous or inaccurate responses when used in non-English or racially diverse populations (Omiye et al., 2023). Furthermore, several language models are optimized for English tokenization, which may result in lower performance when interacting with languages such as Spanish or Portuguese. This has direct implications for populations in LMICs, where cultural, linguistic, and socioeconomic factors deeply influence the experience and expression of mental health issues. Previous studies show that AI models can produce biased, culturally inappropriate, or clinically incorrect responses, reinforcing existing inequities in marginalized groups, such as Afro-descendant populations or those in poverty (Cross et al., 2024; Omiye et al., 2023). For example, an AI-generated response might misinterpret local idioms associated with emotional suffering or provide recommendations that overlook the structural realities of unequal access to healthcare services. This is especially relevant in LMICs, where the social and cultural determinants of mental health are complex. These biases can amplify disparities and affect the quality of support received (Ahluwalia et al., 2025). This relates to the principle of health equity, reminding us that no technology is inherently neutral or universal.

The expansion of generative AI use in mental health also raises questions regarding legal liability. Who is responsible for harm derived from a potentially dangerous, incomplete, or erroneous recommendation? This debate has intensified following recent lawsuits against AI providers for adverse outcomes associated with their responses. For example, documented harms include responses that promote discrimination, hate speech, or exclusion; harms arising from misinformation or malicious uses; harms related to human-computer interaction; and environmental or socioeconomic harms (Weidinger et al., 2022).

At the clinical level, available evidence is limited. Although some exploratory studies suggest that AI-based chatbots can support psychological interventions (Baek et al., 2025; Li et al., 2023; Vaidyam et al., 2019; Villarreal-Zegarra et al., 2024), few of these generative systems have been evaluated through large, rigorous clinical trials that support their efficacy or safety, as most sample sizes are small and follow-up times are short (Li et al., 2023). Furthermore, the structural "black box" problem of generative AI models hinders a clear understanding of how responses are generated, making it challenging to assess risks, validate recommendations, and ensure therapeutic coherence (Ethical AI governance group, 2023).

Without robust data derived from clinical studies and without clear legislative mechanisms for clinical and regulatory oversight, the integration of generative AI into mental health practices may pose more risk than benefit. This is particularly relevant because many LMICs lack governance frameworks adapted to generative AI (Smart & Montori, 2025; Stanford Center for Digital Health, 2025). Specifically in Latin America, the absence of specific legislation on AI in healthcare, combined with the lack of local clinical trials or studies that include Latin American populations, creates a context where the risks outweigh the potential benefits.

This situation demands coordinated actions among multiple actors. In this letter to the editor, we call upon researchers, AI technology developers, and policymakers. First, research teams must prioritize validation studies in real-world contexts, using heterogeneous samples that represent underrepresented populations, including clinical trials, evaluations of cultural bias, and analyses of unintended effects. This will ensure that the performance of AI-based applications in healthcare is sufficiently robust and minimizes biases that can generate inequities. Second, developers must incorporate ethical principles into the design of these applications, guaranteeing transparency, traceability of algorithmic decisions, and safety safeguards in crisis scenarios. Third, regulators and public health authorities must develop specific guidelines for generative AI technologies that address privacy, equity, civil liability, and clinical validity. This includes adapting frameworks, such as those of the WHO on digital governance, to the realities of low- and middle-income countries.

We believe that low- and middle-income countries have the opportunity to anticipate risks and establish an ethical and regulatory framework that protects users, particularly those seeking emotional and mental health support in vulnerable situations. Without these actions, the promise of generative AI in health services could become a new form of digital and health inequity.

ORCID

Leonardo Rojas-Mezarina: https://orcid.org/0000-0003-0293-7107
David Villarreal-Zegarra: https://orcid.org/0000-0002-2222-4764

AUTHORS' CONTRIBUTION

Leonardo Rojas-Mezarina: Conceptualization, investigation, writing - original draft, and approval of the final version.
David Villarreal-Zegarra: Review, investigation, writing - original draft, and approval of the final version.

FUNDING

This paper has been self-financed.

CONFLICT OF INTEREST

The authors declare that there were no conflicts of interest in the collection, analysis, or writing of the manuscript.

ACKNOWLEDGMENTS

Not applicable.

REVIEW PROCESS

This study has been reviewed by external peers in double-blind mode. The editor in charge was Renzo Rivera. The review process is included as supplementary material 1.

DATA AVAILABILITY STATEMENT

Not applicable.

DECLARATION OF THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE

We used DeepL to translate specific sections of the manuscript to Spanish and Grammarly to improve the wording of certain sections. The final version of the manuscript was reviewed and approved by all authors.

DISCLAIMER

The authors are responsible for all statements made in this article.

REFERENCES

Ahluwalia, M., Sehgal, S., Lee, G., Agu, E., & Kpodonu, J. (2025). Disparities in Artificial Intelligence-Based Tools Among Diverse Minority Populations: Biases, Barriers, and Solutions. *JACC. Advances*, *4*(5), 101742. https://doi.org/10.1016/j.jacadv.2025.101742

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, *183*(6), 589-596. https://doi.org/10.1001/jamainternmed.2023.1838

Baek, G., Cha, C., & Han, J.-H. (2025). AI Chatbots for Psychological Health for Health Professionals: Scoping Review. *JMIR Human Factors*, *12*, e67682. https://doi.org/10.2196/67682

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P. (2023). *The Foundation Model Transparency Index* (No. arXiv:2310.12941). arXiv. https://doi.org/10.48550/arXiv.2310.12941

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). *How People Use ChatGPT* (Working Paper No. 34255). National Bureau of Economic Research. https://doi.org/10.3386/w34255

Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, *3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

Ethical AI governance group. (2023). *BEYOND THE BLACK BOX: Shaping a Responsible AI Landscape*. KPMG.

Hussain, S. A., Bresnahan, M., & Zhuang, J. (2025). Can artificial intelligence revolutionize healthcare in the Global South? A scoping review of opportunities and challenges. *Digital Health*, *11*, 20552076251348024. https://doi.org/10.1177/20552076251348024

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

Mandal, A., Chakraborty, T., & Gurevych, I. (2025). *Towards Privacy-aware Mental Health AI Models: Advances, Challenges, and Opportunities* (No. arXiv:2502.00451). arXiv. https://doi.org/10.48550/arXiv.2502.00451

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Medicine*, *6*(1), 120. https://doi.org/10.1038/s41746-023-00873-0

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *26*(4), 2141-2168. https://doi.org/10.1007/s11948-019-00165-5

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. *Npj Digital Medicine*, *6*(1), 195. https://doi.org/10.1038/s41746-023-00939-z

Smart, S., & Montori, V. M. (2025, abril 23). *Peru's AI Regulatory Boom: Quantity Without Depth?* https://www.hks.harvard.edu/centers/carr-ryan/our-work/carr-ryan-commentary/perus-ai-regulatory-boom-quantity-without-depth

Stanford Center for Digital Health. (2025). *AI for Health in Low- and MiddleIncome Countries*. CDH.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930-1940. https://doi.org/10.1038/s41591-023-02448-8

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie*, *64*(7), 456-464. https://doi.org/10.1177/0706743719828977

Villarreal-Zegarra, D., Reategui-Rivera, C. M., García-Serna, J., Quispe-Callo, G., Lázaro-Cruz, G., Centeno-Terrazas, G., Galvez-Arevalo, R., Escobar-Agreda, S., Dominguez-Rodriguez, A., & Finkelstein, J. (2024). Self-Administered Interventions Based on Natural Language Processing Models for Reducing Depressive and Anxiety Symptoms: Systematic Review and Meta-Analysis. *JMIR Mental Health*, *11*, e59560. https://doi.org/10.2196/59560

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., Granger, D., Birse, M., Branson, R., Moons, K. G. M., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ (Clinical Research Ed.)*, *368*, l6927. https://doi.org/10.1136/bmj.l6927

Wang, Y.-F., Li, M.-D., Wang, S.-H., Fang, Y., Sun, J., Lu, L., & Yan, W. (2025). Large language models in clinical psychiatry: Applications and optimization strategies. *World Journal of Psychiatry*, *15*(11), 108199. https://doi.org/10.5498/wjp.v15.i11.108199

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of Risks posed by Language Models.