Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Designing Ethical Artificial Intelligence-Driven Mental Health Applications: *A Cross-Contextual Framework Informed by Developers, Digital Ethics Experts, and Digital Health Specialists in Switzerland and Nigeria*

Lorenta Ojo

lorenta.ojo@ibme.uzh.ch

University of Zurich https://orcid.org/0009-0004-2978-4467

Prof. Markus Christen

University of Zurich

Research Article

Keywords: artificial intelligence, mental health, digital health, biomedical ethics, algorithmic fairness, privacy, qualitative research, low- and middle-income countries, ethical framework

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10220592/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Artificial intelligence (AI)-driven mental health applications are proliferating rapidly across high-income and low- and middle-income countries (LMICs), yet the ethical governance of their design and deployment remains inadequately operationalised. Existing frameworks are largely principle-based and fail to address how ethical obligations are understood and implemented by those building and evaluating these systems.

Objective

To develop an empirically grounded, context-sensitive ethical design framework for AI-driven mental health applications through qualitative stakeholder research.

Methods

Ten semi-structured interviews were conducted with three stakeholder groups: mental health app developers, digital ethics experts, and digital health specialists in Switzerland and Nigeria. Data were analysed using iterative thematic analysis guided by six conceptual domains: transparency, privacy, algorithmic fairness, safety, accountability, and contextual adaptation. Dual ethical approval was obtained from the University of Zurich Ethics Committee (Review No. MeF-Ethik-2026-10) and IAMRAT, Nigeria (Protocol No. UI/EC/26/0187).

Results

Eight themes emerged. AI was conditionally accepted as a bounded, supportive tool rather than an autonomous clinical authority. Human oversight was the strongest cross-stakeholder consensus. Privacy was universally prioritised but governance maturity varied. Fairness required both technical performance parity and contextual cultural fit. Accountability remained fragmented in practice. Nigeria-based participants emphasised scalability, multilingual access, and infrastructure constraints, whereas Swiss participants prioritised regulatory compliance, evidentiary standards, and validation.

Conclusions

A 'core-plus-context' model is proposed: a universal ethical foundation comprising privacy, fairness, transparency, safety, human oversight, and accountability, coupled with context-sensitive implementation. An eight-pillar ethical design framework is presented to guide developers, clinicians, and policymakers toward responsible AI-driven mental health innovation across diverse global settings.

1. Background

Artificial intelligence (AI) is increasingly embedded in mental health applications, transforming how symptoms are assessed, risks are evaluated, and clinical or near-clinical support is provided to users and practitioners (Torous, 2025; Wajid, 2025). These technologies hold genuine promise: they expand access, enable personalisation, and offer scalable support in settings where specialist mental health services are scarce. Yet they also introduce substantial ethical risks. AI-driven mental health applications actively interpret sensitive user data, infer internal psychological states, and influence decisions with potentially serious clinical and social consequences.

The literature documents several concerning patterns. AI systems used in mental health contexts may exhibit biased performance across demographic subgroups, fail to generalise beyond training populations, and respond inadequately to crisis signals such as suicidal ideation (Grabb et al., 2024; Adler et al., 2024). Large language model (LLM)-based conversational agents, now widely deployed in wellness and quasi-clinical contexts, operate as 'black boxes', generating outputs that appear authoritative while concealing uncertainty, limitations, and error rates (Blease, 2024). These properties undermine informed consent and increase the risk of inappropriate reliance by both users and clinicians.

These challenges intensify when applications are deployed across diverse socio-economic and healthcare contexts. High-income countries such as Switzerland typically benefit from stronger regulatory frameworks, established data protection laws, and closer integration of digital tools within clinical governance structures. Conversely, low- and middle-income countries (LMICs) such as Nigeria face distinct pressures: limited specialist mental health infrastructure, variable digital literacy, linguistic diversity, and the absence of locally representative training data. AI systems designed primarily for high-income contexts may generate new inequities or cause harm when exported without adequate adaptation or ethical safeguards (Olawade, 2024; Naslund et al., 2017).

Despite rapid growth in the field, a persistent gap exists between high-level ethical principles and their practical operationalisation in development practice. Most existing literature offers either conceptual analysis or technical performance evaluation, with limited empirical attention to how ethical principles are understood and implemented by developers, ethics experts, and digital health specialists, especially across both HIC and LMIC settings (Cruz-Gonzalez, 2025; Putica et al., 2025). The present study addresses this gap directly.

2. Methods

2.1 Study Design

A qualitative, stakeholder-informed design was employed. The study settings were Switzerland and Nigeria, selected to enable cross-contextual comparison between a high-income regulatory environment and an LMIC context characterised by rapid AI adoption and significant service scarcity.

2.2 Ethical Clearance

Ethical approval

was obtained from the University of Zurich Ethics Committee (Review No. MeF-Ethik-2026-10) and the Institute of Advanced Medical Research and Training (IAMRAT), University of Ibadan, Nigeria (Protocol No. UI/EC/26/0187). All participants provided voluntary, informed consent. Data were anonymised at analysis.

2.3 Participants and Sampling

Participants were recruited through purposive and availability sampling across three groups: (1) mental health app developers; (2) digital ethics stakeholders with expertise in AI governance, biomedical ethics, or technology policy; and (3) digital health specialists with clinical or implementation experience in digital mental health. Ten participants were interviewed in total.

2.4 Data Collection

In-person and remote semi-structured interviews were conducted (30–60 minutes each). Three tailored interview guides were developed, each encompassing up to 20 questions covering: AI functionality and scope; privacy and data governance; bias detection and mitigation; transparency and consent mechanisms; crisis handling and safety protocols; accountability structures; and cross-contextual ethical priorities. Interviews were audio-recorded with consent and transcribed verbatim.

2.5 Data Analysis

Thematic analysis was conducted using an iterative, inductive-deductive coding approach. A structured codebook was developed from the transcripts, comprising 15 codes and associated subthemes, clustered into four broad thematic domains: ethical foundations, safety and legitimacy, context and inclusion, and governance in practice. Analytical memos documented interpretive reasoning. Data from Swiss and Nigerian participants were systematically compared to identify convergent and divergent findings.

3. Results

Thematic analysis generated eight major themes. Across the dataset, AI was not rejected as inappropriate for mental health use. Rather, participants articulated conditional acceptance: AI may offer meaningful benefits in access, responsiveness, scalability, and personalisation, provided it is bounded in function, ethically governed, and supported by human oversight. The following summarises the key findings.

3.1 Theme 1: AI as Bounded Support, Not Autonomous Authority

Participants described AI as useful in supportive, triage-oriented, and engagement-focused roles. Nigerian developers reported broader functional deployment including multilingual conversational interfaces, mood-based recommendation engines, and crisis escalation systems. Swiss participants described more constrained AI roles, often preferring rule-based or limited augmentation over autonomous decision support.

An important observation is that ethical significance in mental health AI does not depend solely on whether a system performs formal diagnosis. Even 'wellness' or 'supportive' tools can shape emotional decision-making, delay care-seeking, and influence crisis response. The interviews revealed a risk of unintended use-case expansion, termed 'spilling over', in which tools designed for general purposes drift into mental health functions without adequate anticipatory governance.

3.2 Theme 2: Human Oversight as the Practical Ethical Minimum

The most robust cross-stakeholder consensus concerned human-in-the-loop oversight. All groups affirmed that AI mental health applications require meaningful human involvement, particularly where elevated risk is present. However, 'human-in-the-loop' was operationalised differently: from continuous clinician authority to reactive escalation triggered by specific risk thresholds. The framework therefore treats oversight as an engineered workflow specifying who intervenes, when, with what competence, and under what conditions, rather than a symbolic safeguard.

3.3 Theme 3: Privacy Is Universally Valued, but Governance Maturity Is Uneven

Privacy and data protection were among the most frequently raised concerns. Ethics stakeholders framed privacy in terms of trust, autonomy, and power asymmetries. Developers demonstrated awareness of obligations and referenced GDPR, NDPR, and HIPAA. However, governance specificity varied: in several cases, clarity was absent regarding retention periods, reuse for model training, and user rights to contest or withdraw data. The analysis supports a distinction between privacy awareness, which was high, and privacy governance maturity, which was uneven.

3.4 Theme 4: Technical Fairness and Contextual Fairness Must Both Be Addressed

Bias and fairness were treated as serious concerns by all groups. Developers described diversity in testing and local language adaptation. However, limited evidence appeared of formal subgroup audits or public performance reporting. Ethics and digital health stakeholders extended fairness beyond statistical performance to encompass language barriers, cultural assumptions, literacy gaps, and the risks of importing models from high-income settings without local adaptation. The study introduces a distinction between technical fairness (measurable performance parity) and contextual fairness (the social and

cultural appropriateness of the system in its environment of use), as one of its central conceptual contributions.

3.5 Theme 5: Transparency and Consent Must Be Layered, Ongoing, and Meaningful

Ethics stakeholders consistently challenged thin, one-off disclosure practices. Meaningful consent, in their view, is an ongoing communicative process, modular by function, accessible in plain language, and open to revision. Developers described pop-ups and user guidance documents, which often fell short of the layered consent models envisioned. The study proposes 'calibrated legibility': sufficient explanation for users and clinicians to judge what the AI can and cannot do, and when human care is necessary.

3.6 Theme 6: Clinical Validity Is the Dividing Line

Digital health stakeholders consistently imposed a stronger evidence threshold than developers. They challenged the assumption that innovation, scalability, or engagement alone are sufficient grounds for deployment. This created the clearest cross-group tension in the dataset: an innovation logic (focus on reach and function) versus a clinical assurance logic (focus on evidence and safety). The ethical framework must specify what level of evidence is required at each level of AI decision influence.

3.7 Theme 7: Accountability Is Distributed Across the Lifecycle

Participants agreed that AI systems cannot themselves be accountable; responsibility lies with humans and institutions. However, uncertainty remained about responsibility allocation when harm occurs. Ethics stakeholders called for explicit responsibility structures, audit trails, and accessible redress mechanisms. The dataset supports the articulation of distributed accountability: developers, concept designers, clinicians, and institutions share responsibility. The 'wellness/not-quite-clinical' grey zone is especially dangerous, as applications may avoid medical device classification while mediating consequential quasi-clinical decisions.

3.8 Theme 8: Context Changes Implementation Priorities, Not the Ethical Core

Nigerian participants foregrounded infrastructure limitations, multilingual requirements, affordability, variable digital literacy, and the moral urgency of extending access. Swiss participants focused on conditions for responsible integration: validation, governance, explainability, and evidence thresholds. These differences represent divergent implementation pressures, not divergent ethical values. The frameworks most useful for this field must therefore be universal in values but context-sensitive in implementation.

4. Discussion

The eight themes identified correspond closely to the domains structuring the study, confirming that the concerns identified empirically are well-grounded in current international debates on AI ethics in health (WHO, 2021). The study extends those debates by showing how ethical principles are interpreted and operationalised differently across stakeholder groups and national contexts.

The centrality of human oversight confirms the literature's warnings about AI reliability and safety in mental health contexts (Gamble, 2020; APA, 2024). The study's contribution lies in revealing that 'human-in-the-loop' is invoked as consensus but operationalised in substantially different ways. Future guidance should define the timing, scope, authority, and competence of human review with greater specificity.

The distinction between technical fairness and contextual fairness is a significant conceptual contribution. Existing literature correctly identifies algorithmic bias as a concern (Adler et al., 2024), but this study demonstrates that fairness in AI mental health tools must additionally address cultural and social appropriateness, not merely statistical performance parity. This is particularly important in LMIC contexts where systems trained in high-income environments may fail to capture local idioms of distress, cultural norms, and contextually appropriate help-seeking pathways (Naslund et al., 2017; Chakrabarti, 2024).

The ethics-implementation gap, a major premise of this study, was strongly validated. Developers framed ethics in operational terms, while ethics stakeholders raised structural concerns about autonomy, justice, and lifecycle governance. This divergence reflects not a lack of ethical concern among developers, but insufficient access to practical, context-sensitive frameworks that translate normative principles into actionable design requirements. The eight-pillar framework presented here addresses this directly.

Cross-contextual comparison yielded a clear finding: ethical principles are largely shared across Switzerland and Nigeria, but their practical implementation differs. This supports a 'core-plus-context' model: a common ethical floor, a context-sensitive implementation layer, and iterative review mechanisms. A single static global standard is both insufficient and potentially inequitable.

5. The Eight-Pillar Ethical Design Framework

The following framework synthesises the empirical findings and the wider literature. It is structured around a core-plus-context model. The hallmark of this study is the convergence of eight themes, eight key findings, and eight pillars.

Pillar	Core Principle	Context-Sensitive Considerations	Interpretation from Data
1. Purpose and Risk Definition	Explicit definition of app function, foreseeable harms, and required evidence level.	LMIC: access vs. safety; HIC: regulatory alignment.	Many ethical tensions begin when app claims are ambiguous.
2. Proportionate AI Use	Least risky architecture sufficient for the intended function.	LMIC: constrained infrastructure; HIC: evidence thresholds.	Swiss participants were notably cautious; developers showed broader use cases.
3. Data Stewardship and Contextual Privacy	Minimisation, storage transparency, retention limits, user withdrawal rights, model use disclosure.	LMIC: jurisdictional gaps; HIC: GDPR/Swiss compliance.	All groups stressed privacy but operational maturity varied.
4. Fairness and Cultural Fit	Technical subgroup and contextual cultural usability audits across language, literacy, age, gender, and local norms.	LMIC: local validation essential; HIC: subgroup reporting standards.	Bias recognised everywhere; contextual fairness especially important.
5. Human Oversight and Escalation	Role-defined, timely, competent oversight; explicit emergency pathways.	LMIC: non-specialist escalation; HIC: clinician governance.	Strongest consensus across the dataset.
6. Meaningful Transparency and Consent	Layered, ongoing, modular, plain-language consent; training data provenance; AI disclosure.	Universal; adapted to literacy level and delivery channel.	Experts asked for more than standard privacy notices.
7. Accountability and Auditability	Distributed responsibility, version control, incident reporting, model cards, user redress.	LMIC: lightweight audit structures; HIC: regulatory documentation.	Accountability accepted in principle but diffuse in practice.
8. Lifecycle Evaluation	Post-deployment monitoring, adverse event surveillance, drift assessment, crisis misses, user trust outcomes.	Both contexts: proportionate to risk and user vulnerability.	Participants explicitly called for monitoring.

6. Conclusions

This study demonstrates that ethical AI in mental health cannot be reduced to the addition of safety features, the production of privacy policies, or the declaration of abstract principles. Ethical acceptability requires a full, practical, context-aware governance model that defines the application's clinical role, limits AI authority, protects vulnerable users, ensures meaningful consent, validates performance technically and culturally, distributes accountability, and maintains post-deployment monitoring.

The strongest shared finding is the centrality of human oversight. Privacy, fairness, and transparency must be understood in deeper ways than product rhetoric often suggests. Privacy requires governance, not only security; fairness requires contextual as well as technical sensitivity; transparency requires explanation that helps users and professionals calibrate trust.

The core-plus-context model proposed here offers a structured response to the inadequacy of both single static global standards and infinitely flexible local approaches. This framework, grounded in empirical evidence from developers, ethics experts, and digital health specialists across two national settings, constitutes a practical and replicable contribution to responsible AI-driven mental health innovation globally.

7. Recommendations

7.1 For Developers

Adopt the eight-pillar framework as the minimum governance architecture: document the intended clinical role; select AI architectures proportionate to the evidence base; engineer privacy into the design; test bias across technical and contextual dimensions; define human oversight workflows explicitly; implement layered consent mechanisms; document accountability across the lifecycle; and conduct continuous post-deployment monitoring.

7.2 For Clinicians and Digital Health Stakeholders

Evaluate AI mental health applications according to the degree of influence they exert over care decisions. Maintain interpretive authority and resist treating AI outputs as self-validating. Actively co-produce ethical standards with developers and ethics experts.

7.3 For Policymakers and Regulators

Adopt risk-sensitive governance frameworks calibrated to whether an application handles crises, influences clinical pathways, or uses opaque AI systems. Support locally representative data infrastructure, multilingual accessibility, and governance mechanisms adapted to local realities in lower-resource settings.

7.4 Limitations and Future Research

This study's qualitative design and limited sample size constrain generalisability. Swiss perspectives are stronger in ethics and digital health roles, while Nigerian interviews are more developer-heavy. Future research should expand the empirical base by including policymakers, patients, caregivers, and additional LMIC voices. Prospective testing of the framework through developer workshops, expert consensus exercises, and pilot implementation studies is recommended.

Declarations

Ethical Approval

s: University of Zurich Ethics Committee (Review No. MeF-Ethik-2026-10); IAMRAT Nigeria (Protocol No. UI/EC/26/0187)

Ethics Approval and Consent to Participate

The study was conducted in accordance with the Declaration of Helsinki. Ethical clearance was obtained from two institutions. Informed consent was obtained from participants. All data were anonymised, transcribed, and stored securely. Participation was voluntary with no consequences for withdrawal.

Consent for Publication

Not applicable. No individual personal data is published in this article.

Competing Interests

The author declares no competing interests.

Funding

This project was not funded. It was carried out during the author's Stehr Boldt Fellowship at the UZH from 2025–2026.

Author Contributions

All authors contributed substantially to this manuscript. The lead author conceptualized the study, developed the core framework for the AI-driven mental health app design, and led the writing of the manuscript. The supervisors, as co-authors, contributed to the study design, provided critical intellectual input, supervised the methodological direction, and revised the manuscript for important intellectual content. All authors reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Acknowledgements

The primary author gratefully acknowledges the Institute of Biomedical Ethics, University of Zurich and its director for being selected as the sole African Fellow in the 2025 cohort. She would like to sincerely thank her supervisor, Prof. Markus Christen for his invaluable guidance, critical feedback, and continuous academic support throughout the development of this work. His expertise in biomedical ethics, digital health innovation and research methodology significantly strengthened the conceptualization and

refinement of this article. She also extends appreciation to IAMRAT Nigeria; peers and colleagues at the Zurich institute and members of Digital Society Initiative (DSI) who provided insights during the formative stages of the design process. Gratitude is given to the ten amazing experts who were particularly instrumental in shaping the ethical and user-centered considerations of AI-driven mental health applications. Finally, special thanks go to Chinwendu Egwuagu and Franklin Ndupu for their boundless support without which the Fellowship wouldn't have been easy to undertake.

Availability of Data and Materials

The datasets used and analysed in the current study are available from the corresponding author on reasonable request.

References

1. Adler DA, Wang F, Mohr DC, Torous J (2024) Measuring algorithmic bias to analyse the reliability of AI tools that predict depression risk using smartphone-sensed behavioural data. npj Digit Med 7(1). https://doi.org/10.1038/s41746-024-01159-9
2. American Psychological Association (2024) Artificial intelligence in mental health care. https://www.apa.org/about/policy/artificial-intelligence-mental-health.pdf
3. Blease C (2024) Generative artificial intelligence in mental healthcare: An ethical evaluation. J Med Ethics 50(3):183–189. https://doi.org/10.1136/jme-2023-109388
4. Chakrabarti S (2024) Digital psychiatry in low- and middle-income countries: New developments and the way forward. World J Psychiatry 14(3):413–427. https://doi.org/10.5498/wjp.v14.i3.413
5. Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Lohmann J, Robberstad B, Neuhann F (2022) Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: A systematic scoping review. npj Digit Med 5(1):162. https://doi.org/10.1038/s41746-022-00700-y
6. Cruz-Gonzalez P (2025) Artificial intelligence in mental health care: A systematic review. [Manuscript under review.]
7. Gamble A (2020) Artificial intelligence and mobile apps for mental healthcare: A social informatics perspective. Aslib J Inform Manage 72(4):509–523. https://doi.org/10.1108/AJIM-11-2019-0316
8. Grabb D, Lamparth M, Vasan N (2024) Risks from language models for automated mental healthcare: Ethics and structure for implementation. arXiv preprint arXiv :240115768
9. Linardon J, Fuller-Tyszkiewicz M, Firth J, Torous J, Messer M (2024) Systematic review and meta-analysis of adverse events in clinical trials of mental health apps. npj Digit Med 7(1). https://doi.org/10.1038/s41746-024-01388-y
10. Naslund JA, Aschbrenner KA, Araya R, Marsch LA, Unutzer J, Patel V, Bartels SJ (2017) Digital technology for treating and preventing mental disorders in low-income and middle-income countries. Lancet Psychiatry 4(6):486–500. https://doi.org/10.1016/S2215-0366(17)30096-2

11. Olawade DB (2024) AI in mental health: Considerations for low- and middle-income countries. [Manuscript under review.]
12. Putica A, Harrison M, Bhavani N, Torous J, Faurholt-Jepsen M, Moran P (2025) Ethical decision-making for AI in mental health: The IEACP framework. Comput Psychiatry 9(1):1–18. https://doi.org/10.1162/cpsy_a_00102
13. Radanliev P (2025) Privacy, ethics, transparency, and accountability in AI systems for wearable devices. Front Digit Health 7., Article 1431246. https://doi.org/10.3389/fdgth.2025.1431246
14. Tavory T (2024) Regulating AI in mental health: An ethics of care perspective. JMIR Mental Health 11:e56543. https://doi.org/10.2196/56543
15. Torous J (2025) The evolving field of digital mental health. World Psychiatry 24(1):12–19. https://doi.org/10.1002/wps.21244
16. Wajid A (2025) Applications of artificial intelligence in mental health. In: Chandra P, Bhargava R (eds) Digital innovations in global health. Springer, pp 241–268
17. Wang X (2025) Generative AI in mental health: Applications and ethics. JMIR Mental Health 12:e59011. https://doi.org/10.2196/59011
18. World Health Organization (2021) Ethics and governance of artificial intelligence for health. WHO. https://www.who.int/publications/i/item/9789240029200

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Stunningpic.png