How Human-AI Collaboration Illuminates Mental Health Care

Ratna Kandala[1], **Akshata Kishore Moharir**[2]
[1]University of Kansas, KS, USA
[2]Independent Researcher, Oregon, USA

n038k926@ku.edu, akshatan@terpmail.umd.edu

Abstract

This paper presents a comprehensive framework for responsible and explainable AI (XAI) in mental health screening (MHS), bridging the gap between technical XAI tools and clinical requirements.

Background – What Is Known?

AI has shown growing potential in mental health care, from detecting depression through language patterns to assisting in clinical diagnostics. With the rise of large language models (LLMs), interest in AI-powered tools has intensified across public and clinical domains (Nori et al. 2023). At the same time, the mental health domain presents unique ethical challenges due to the sensitivity of data, cultural variability in diagnosis, and the high stakes of misinterpretation (Mittelstadt et al. 2016; Jobin, Ienca, and Vayena 2019).

Problem – What Is Unknown or Underdeveloped?

Despite technical advances in XAI, current models remain poorly aligned with the realities of mental health practice. Common tools like LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) aim to make black-box models interpretable but often lack clinical relevance or cultural sensitivity (Lipton 2016; Rudin 2019; Doshi-Velez and Kim 2017). Moreover, there is a gap in actionable frameworks that integrate these models responsibly into clinical workflows (Khan et al. 2025). Issues of trust, inclusion, and long-term impact remain underexplored, and existing benchmarks fail to evaluate fairness, adaptability, or real-world utility. This synthesis revealed that no single method is universally sufficient, motivating a framework that integrates complementary strengths rather than privileging one approach. The four pillars thus emerge directly from the gaps exposed by the synthesis rather than being chosen arbitrarily.

Contribution – What This Work Contributes?

This work bridges the gap between technical XAI tools and the nuanced requirements of mental healthcare. We provide a multi-pronged strategy for aligning XAI systems with clinical and ethical priorities:

- A systematic synthesis of XAI methods tailored to mental health, highlighting case-based reasoning, Chain-Of-Thought (CoT) prompting, and retrieval-augmented generation (RAG).
- A strategic blueprint for responsible deployment grounded in participatory co-design, human-centered evaluation metrics, and the proposal of a "living benchmark" for mental health systems.
- A call to reframe AI alignment for mental health beyond "helpful, honest, and harmless" toward systems that are empathetic, culturally aware, and accountable (Gabriel 2020; Crawford 2021).

Proposed Framework

To address critical gaps in responsible AI deployment for mental health, we outline a four-part framework. The systematic synthesis of XAI methods serves as the empirical foundation for the four-part framework, with findings on LLM-based explanations, dataset gaps, and static benchmarks directly shaping the human-centered metrics, benchmarking for inclusion, and living benchmark pillars respectively. Together, the synthesis and framework operate as diagnosis and prescription, one maps the landscape of current XAI capabilities and limitations, the other provides a roadmap for responsible deployment:

- **Participatory Co-Design:** Involve clinicians, patients, and marginalized communities in system development (Birhane 2021).
- **Human-Centered Metrics:** Prioritize comprehensibility, trust calibration, and long-term impact over mere accuracy (Joyce et al. 2023).
- **Benchmarking For Inclusion:** Address the lack of representative datasets and culturally valid evaluation tools (Gebru et al. 2018; Zhao et al. 2023).
- **Living Benchmark:** Introduce a dynamic benchmark that evolves with real-world data and integrates fairness and robustness.

We acknowledge that human-AI collaboration in mental health is an iterative, trust-mediated process in which

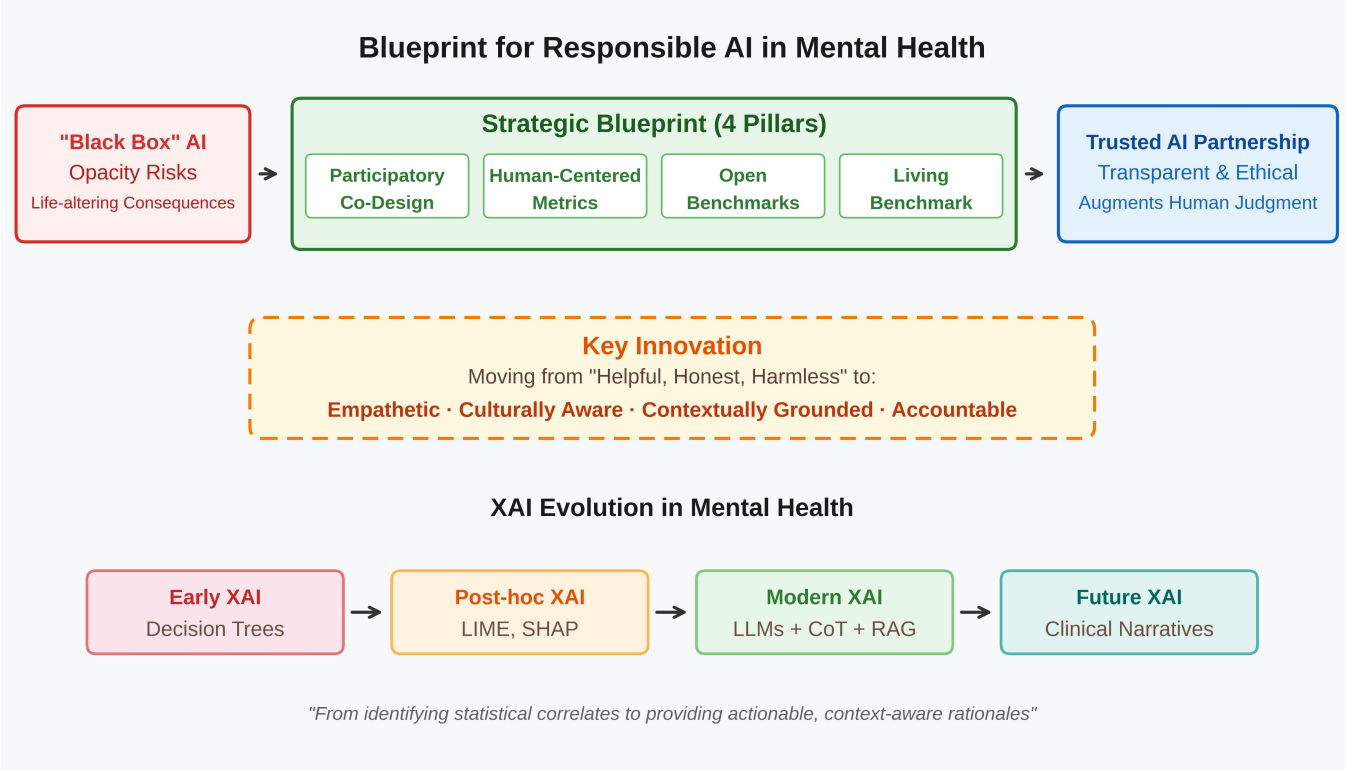

Figure 1: Strategic blueprint for a responsible Human-AI alliance in mental healthcare.

clinicians interact with AI outputs in a decision loop, interrogating explanations, overriding recommendations, and feeding corrections back into system refinement rather than passively accepting predictions (Birhane 2021). In this regard, methods like case-based reasoning, CoT prompting, and RAG each support distinct aspects of this collaboration, with the shared goal of augmenting rather than replacing clinical judgment while preserving clinical agency and accountability (Ahmed et al. 2025).

Summary of Findings

We trace the evolution of XAI in mental health from early feature importance scores (e.g., identifying keywords like "hopeless" (Eichstaedt et al. 2018)) to systems that generate context-rich, clinically coherent explanations using LLMs (Ahmed et al. 2025). We demonstrate that newer models offer more than interpretability-they contribute to trust-building through explainability that aligns with clinician reasoning and patient understanding.

Remaining Challenges

We identify key open questions guiding future research:

- How can trust in AI be seen as dynamic and socially constructed?
- How can explanations enhance user agency instead of dictating clinical meaning?
- How can we mitigate both short- and long-term harms, from algorithmic bias to over-medicalization?
- How can systems adjust to cultural and epistemological pluralism?

Future Work

To validate the framework empirically, we outline a human subjects study where clinicians evaluate AI-generated explanations for comprehensibility and clinical relevance, while patients assess trust and perceived empathy. Outcomes are measured through trust calibration scores and clinical decision alignment. This study follows participatory evaluation principles with full IRB approval, informed consent, and feedback loops that enable clinician corrections to inform ongoing system refinement.

Conclusion

The future of AI in mental health relies on tools that not only "work" but also earn trust, respect complexity, and enhance human judgment. This work establishes a new generation of mental health AI—technically robust, ethically sound, and aligned with the humanistic principles central to mental health care.

References

Ahmed, A.; Saleem, M.; Alzeen, M.; Birur, B.; Fargason, R. E.; Burk, B. G.; Alhassan, A.; and Al-Garadi, M. A. 2025.

Explainable AI for mental health emergency returns: Integrating LLMs with predictive modeling. arXiv:2502.00025.

Birhane, A. 2021. Algorithmic injustice: a relational ethics approach. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 258–268.

Crawford, K. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. arXiv:1702.08608.

Eichstaedt, J. C.; Smith, R. J.; Merchant, R. M.; Ungar, L. H.; Crutchley, P.; Preoţiuc-Pietro, D.; Asch, D. A.; and Schwartz, H. A. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44): 11203–11208.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3): 411–437.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9): 389–399.

Joyce, D. W.; Kormilitzin, A.; Smith, K. A.; and Cipriani, A. 2023. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, 6(1): 6.

Khan, M. M.; Shah, N.; Shaikh, N.; Thabet, A.; Alrabayah, T.; and Belkhair, S. 2025. Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. *International Journal of Medical Informatics*, 195: 105780.

Lipton, Z. C. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Lundberg, S.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Mittelstadt, B. D.; Allo, P.; Taddeo, M.; Wachter, S.; and Floridi, L. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2).

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rudin, C. 2019. Stop explaining black box models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215.

Zhao, H.; et al. 2023. Explainability for Large Language Models: A Survey. arXiv:2309.01029.