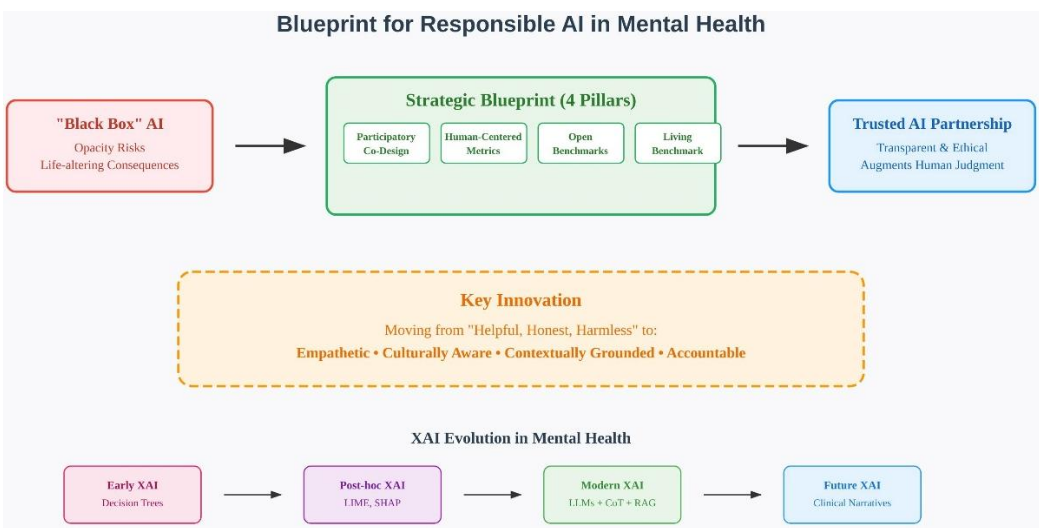
From Algorithm to Alliance: A Blueprint for Responsible and Explainable AI in Mental Health Screening

Background: Explainable AI (XAI) shows promise in mental health, from language-based depression detection to clinical diagnostics. Yet XAI tools like LIME [5] and SHAP [4] remain poorly aligned with clinical realities as they lack cultural sensitivity and actionable deployment frameworks [1]. Critical gaps persist in trust, inclusion, and real-world evaluation, with benchmarks failing to assess fairness and adaptability in high-stakes mental health applications [2,3,6]. This work bridges technical XAI tools with mental healthcare needs through: (a) systematic synthesis of tailored methods (case-based reasoning, Chain-Of-Thought (CoT) prompting, Retrieval Augmented Generation (RAG)), (b) responsible deployment blueprint featuring participatory co-design and "living benchmarks," and (c) reframing AI alignment from helpful/honest/harmless to empathetic, culturally aware, and accountable systems.

Key Findings and Proposed Framework: We trace XAI evolution from keyword detection (e.g., words like “hopeless") to LLM-powered systems generating clinically coherent explanations that build trust by aligning with clinician reasoning and patient understanding. Our four-part approach includes:
(1) participatory co-design with clinicians, patients, and marginalized communities,
(2) human-centered metrics prioritizing comprehensibility over accuracy,
(3) inclusive benchmarking with representative datasets, and
(4) dynamic "living benchmarks" integrating fairness and real-world adaptation.

Open Questions: How can trust be socially constructed? How do explanations support user agency? How do we mitigate algorithmic bias and over-medicalization while adapting to cultural pluralism?

Conclusion: Mental health AI must earn trust, respect complexity, and amplify human judgment. This work establishes foundations for technically robust, ethically sound systems aligned with humanistic mental healthcare principles.



[1] Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on medical challenge problems*. arXiv. https://doi.org/10.48550/arXiv.2303.13375

[2] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

[3] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[4] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates.

[5] Ribeiro, M. T., Singh, S., & 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* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

[6] Khan, M. M., Shah, N., Shaikh, N., Thabet, A., Alrabayah, T., & 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. https://doi.org/10.1016/j.ijmedinf.2024.105780