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

Akshata Kishore Moharir[1], Ratna Kandala
[1]University of Maryland, [2]University of Kansas

Background

AI has shown growing potential in mental health care, from detecting depression through language patterns to assisting in clinical diagnostics
With the rise of LLMs, interest in AI-powered tools has intensified across public and clinical domains
The mental health domain presents unique ethical challenges due to sensitivity of data, cultural variability in diagnosis, and high stakes of misinterpretation

Research Questions

RQ1: How can XAI methods be tailored to mental health screening?
RQ2: What framework supports responsible deployment of AI in mental health?
RQ3: How should AI alignment be reframed beyond 'helpful, honest, and harmless'?
RQ4: What evaluation metrics capture human-centered outcomes?

Contributions

- 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 bench-mark” 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

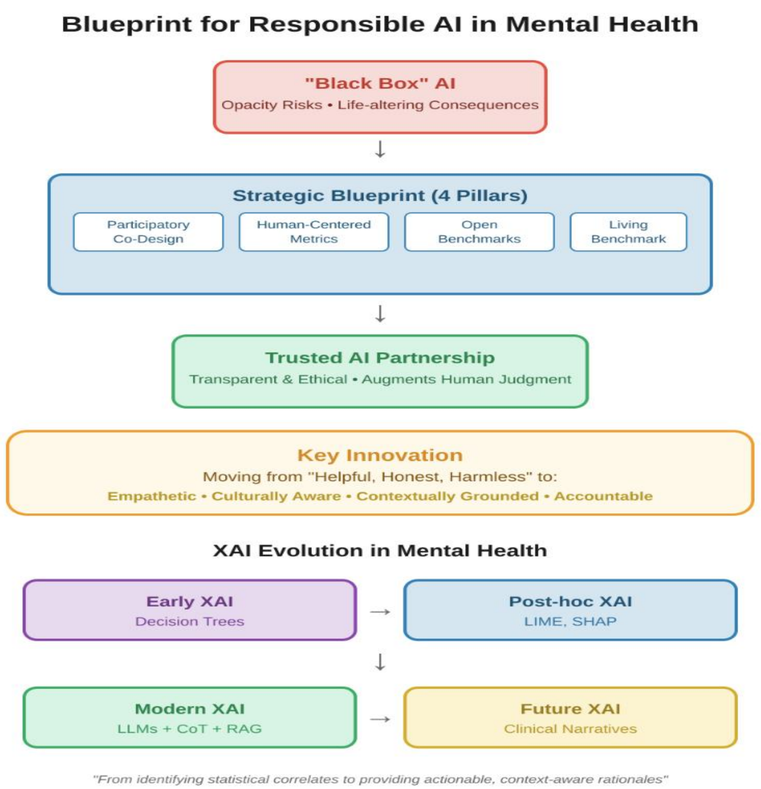
Proposed Framework:

:• Participatory Co-Design: Involve clinicians, patients and marginalized communities in system development.

• Human-Centered Metrics: Prioritize comprehensibility, trust calibration, and long-term impact over mere accuracy

.• Benchmarking for Inclusion: Address the lack of representative datasets and culturally valid evaluation tools

• Living Benchmark: Introduce a dynamic benchmark that evolves with real-world data and integrates fairness and robustness.



Strategy Blueprint for Responsible AI Alliance for Mental HealthCare

Summary of Findings

- Evolution of XAI in mental health traced from early feature importance scores (e.g., "hopeless" keyword detection) to LLM-based context-rich explanations
- Early approaches: Keyword identification
- Current approaches: Clinically coherent explanations using LLMs
- Newer models go beyond interpretability—they build trust by aligning with clinician reasoning and patient understanding

Limitations and Conclusions

We identify open questions that must guide future research:• How can trust in AI be treated as dynamic and socially constructed?
- How can explanations support user agency rather than dictate clinical meaning?
- The future of AI in mental health depends on tools that don’t just “work” but that earn trust, respect complexity, and amplify human judgment. This work lays the foundation fora new generation of mental health AI—technically robust, ethically sound, and aligned with the humanistic principles at the heart of mental health care

References

1. Birhane, A. 2021. Algorithmic injustice: a relational ethics approach. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 258–268
2. Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprintarXiv:1702.08608
3. Zhao, H., et al. 2023. Explainability for large language models: A survey.
4. Mittelstadt, B. D.; Allo, P.; Taddeo, M.; Wachter, S.; andFloridi, L. 2016. The ethics of algorithms: Mapping thedebate. Big Data & Society 3(2).
5. Birhane, A. 2021. Algorithmic injustice: a relational ethicsapproach. In Proceedings of the 2021 ACM conference onfairness, accountability, and transparency, 258–268
6. 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: Inte-grating llms with predictive modeling
7. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.;Wallach, H.; Daum´e III, H.; and Crawford, K. 2018.Datasheets for datasets. arXiv preprint arXiv:1803.09010
8. Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; andHorvitz, E. 2023. Capabilities of GPT-4 on medical chal-lenge problems. arXiv preprint arXiv:2303.13375