



Ethical, Privacy and Fairness Safeguards for AI-Assisted Diagnosis of Depression and Anxiety: A Lifecycle Governance Framework

Mansi*¹, Dr. Grace M. BasanShrieh*²

*¹Phd Scholar, Department of Psychology, William Caery University, Meghalaya, India

*²Assistant Professor and Guide, Department of Psychology, William Caery University, Meghalaya, India

Abstract- Artificial intelligence (AI) tools for detecting depression and anxiety process some of the most sensitive information a person can share: their thoughts, voice, movements and health history. Technical accuracy is therefore necessary but not sufficient for responsible use. This paper, derived from a doctoral study on AI-assisted diagnosis of depression and anxiety, proposes a lifecycle governance framework that translates four ethical principles (autonomy, beneficence and non-maleficence, justice, and explicability) into concrete safeguards at five stages: data collection, model development, validation, deployment and monitoring. Eight principal risks are identified and prioritised on a likelihood–impact matrix: privacy breach, consent failure, subgroup bias, automation bias, dataset shift, opaque output, miscalibration, and stigma or misuse. An analytic fairness demonstration shows how a sensitivity drop from 0.85 to 0.70 in a minority-language subgroup doubles missed cases from 22.5 to 45.0 per 1,000 people screened, even though the overall model appears acceptable. A privacy-preserving data pipeline and a checklist of minimum safeguards are proposed. The framework positions AI as accountable, clinician-supervised decision support and gives researchers, ethics committees and clinicians a practical basis for reviewing mental health AI projects.

Keywords- AI ethics; mental health; depression; anxiety; algorithmic fairness; privacy; informed consent; explainability; clinical governance; automation bias

I. INTRODUCTION

Depression and anxiety are among the most common causes of disability worldwide [1], and many affected people are not identified in routine care [2]. Machine learning (ML) models that analyse questionnaires, health records, language, speech and smartphone data have been proposed to support earlier detection [3]–[5]. Unlike many other clinical AI applications, these tools draw on intimate personal



information and act on conditions that still carry considerable stigma. Errors can lead to missed suicide risk, unnecessary labelling, or unequal treatment of groups that are already under-served.

General frameworks for AI ethics are plentiful, but principles alone do not ensure ethical practice [6]. Researchers and clinicians need concrete, stage-specific safeguards that can be checked. Ethical challenges in clinical ML, including bias, opacity and the shifting of responsibility, have been widely discussed [7], [8], yet few works tailor them to AI-assisted mental health assessment. This paper addresses that need. Its objectives are (i) to translate core ethical principles into safeguards across the AI lifecycle, (ii) to identify and prioritise the principal risks, (iii) to show quantitatively how subgroup bias affects people, and (iv) to propose a privacy-preserving data pipeline and a minimum safeguard checklist.

II. APPROACH

The framework was developed by analysing the ethics, privacy, governance and oversight sections of the doctoral thesis, together with the literature on clinical AI ethics, fairness, explainability and reporting [9], [10]. Four principles from biomedical ethics and AI guidance were adopted: autonomy, beneficence and non-maleficence, justice, and explicability. They were crossed with five lifecycle stages to derive safeguards. Risks were rated by the authors on five-point likelihood and impact scales, assuming no safeguards were in place. A hypothetical scenario was then analysed with standard diagnostic-accuracy formulas to show the effect of subgroup bias.

III. PRINCIPLE-TO-PRACTICE FRAMEWORK

Fig. 1 presents the framework as a matrix. Each cell specifies a safeguard that puts one principle into practice at one stage, so that an ethics committee or project team can check compliance stage by stage.

	Data collection	Model development	Validation	Deployment	Monitoring
Autonomy	Informed, specific, revocable consent	Purpose limitation; no secondary use	Consent to sharing results	Right to decline AI assessment	Right to withdraw data
Beneficence & non-maleficence	Safety protocol for acute risk	Clinical reference standard	Calibration & net benefit	Clinician-in-the-loop	Drift detection; incident reporting
Justice	Representative sampling	Bias-aware features & labels	Subgroup metrics	Equitable access & thresholds	Periodic fairness audit
Explicability	Data provenance records	Interpretable baseline	Faithful explanations	Model facts label	Transparent versioning

Fig. 1. Principle-to-practice matrix: four ethical principles translated into concrete safeguards across five lifecycle stages of an AI-assisted mental health tool.



1. Autonomy and Consent

Consent must be specific to each data type. Agreeing to a questionnaire is not agreeing to voice recording or phone sensing. It must also be revocable and understandable. Participants should know how their data will be used, who can access them, whether AI output will be shared with clinicians, and that they may decline AI assessment without losing access to care. Secondary use of data for purposes not covered by the original consent should be prohibited [11].

2. Beneficence and Non-maleficence

A mental health AI study must include a safety protocol that responds immediately to indications of acute risk, whatever the model outputs. Benefit should be demonstrated against a clinical reference standard, with calibrated probabilities [12] and an assessment of net benefit [13], not accuracy alone. The model must never be the sole basis of a diagnosis.

3. Justice

Justice requires representative sampling, bias-aware choice of features and labels, and reporting of performance for each subgroup [8], [14]. A widely used health algorithm that relied on cost as a proxy for need systematically disadvantaged Black patients [15]. Analogous proxies, such as prior service use, may encode unequal access to mental health care.

4. Explicability

Clinicians and patients should be able to understand what a model does and why it produced a particular output [16]. Interpretable baselines are preferred where they perform comparably [17]. Post-hoc explanations such as SHAP [18] should carry explicit caveats, because they may not faithfully represent model reasoning [19]. A model facts label summarising intended use, population, performance and limitations should accompany any deployment [20].

IV. RISK IDENTIFICATION AND PRIORITISATION

Eight principal risks were identified. Fig. 2 places them on a likelihood–impact matrix, and Table 1 pairs each with safeguards. Subgroup bias (R3) and dataset shift (R5) are rated both likely and of major impact, because models developed in one population frequently degrade silently when used elsewhere [21]. Automation bias (R4), in which clinicians over-rely on a model's output, is well documented in clinical decision support [22]. Privacy breach (R1) is less likely under good controls but has the most severe impact given the sensitivity of mental health data.

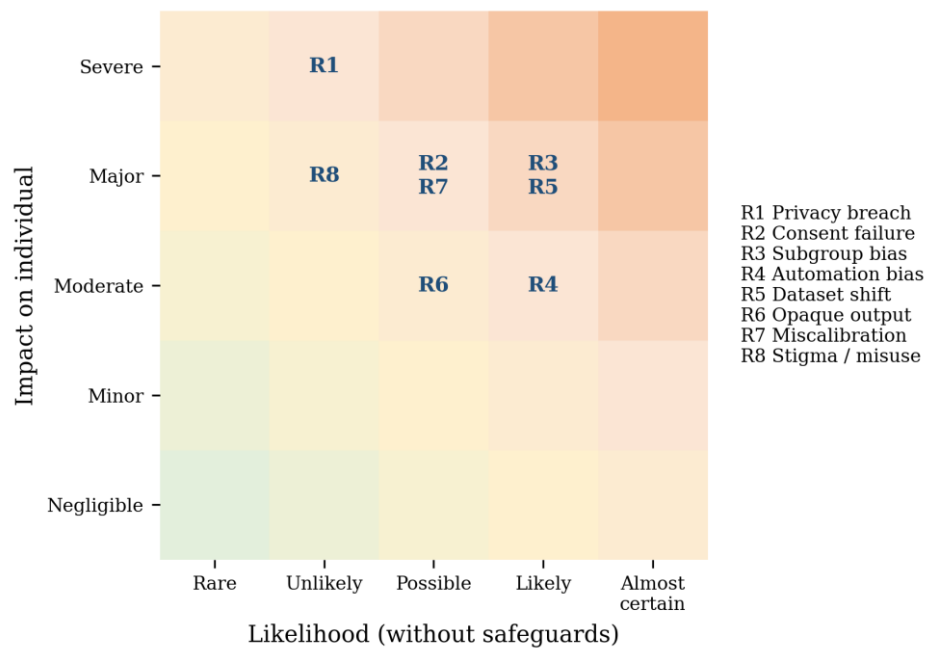


Fig. 2. Likelihood-impact matrix of eight principal risks, assessed without safeguards. Ratings are the authors' qualitative judgement based on the cited literature.

Table 1. Principal risks and minimum safeguards

Risk	Example manifestation	Minimum safeguard
R1 Privacy breach	Exposure of transcripts or diagnoses	Pseudonymisation, encryption, access control, audit trail [11]
R2 Consent failure	Voice or phone data used beyond agreed scope	Modality-specific, revocable consent; purpose limitation
R3 Subgroup bias	Lower sensitivity for minority-language speakers	Representative sampling; subgroup metrics [14]
R4 Automation bias	Clinician accepts incorrect low-risk output	Clinician-in-the-loop; logged overrides [22]
R5 Dataset shift	Performance drop at a new site or on new devices	External validation; drift monitoring [21]
R6 Opaque output	Unexplained flag erodes trust	Interpretable baseline; caveated explanations [19]
R7 Miscalibration	Risk of 0.8 reported where true rate is 0.4	Calibration reporting and recalibration [12]
R8 Stigma / misuse	AI label used by employers or insurers	Use restrictions; no disclosure beyond clinical team



V. FAIRNESS DEMONSTRATION

To show the human consequences of subgroup bias, consider a hypothetical screening model with overall sensitivity and specificity of 0.85. Three subgroups are compared: Group A (majority language, prevalence 15%), Group B (minority language, same prevalence, but sensitivity reduced to 0.70 because of under-representation in training data), and Group C (majority language, prevalence 30%). Positive predictive value (PPV) follows from Bayes' theorem:

$$PPV = (Se \cdot p) / [Se \cdot p + (1 - Sp) \cdot (1 - p)] \quad (1)$$

Missed cases per 1,000 people screened equal $1000 \cdot p \cdot (1 - Se)$, and false alarms equal $1000 \cdot (1 - p) \cdot (1 - Sp)$. As Fig. 3 shows, Group B experiences 45.0 missed cases per 1,000, double the 22.5 in Group A, and PPV falls from 50.0% to 45.2%. Group C shows that differences in prevalence alone change PPV substantially (70.8%), even at identical accuracy. Both effects stay hidden if only aggregate metrics are reported. Subgroup-level reporting of sensitivity, PPV and calibration is therefore a minimum requirement [14], [23].

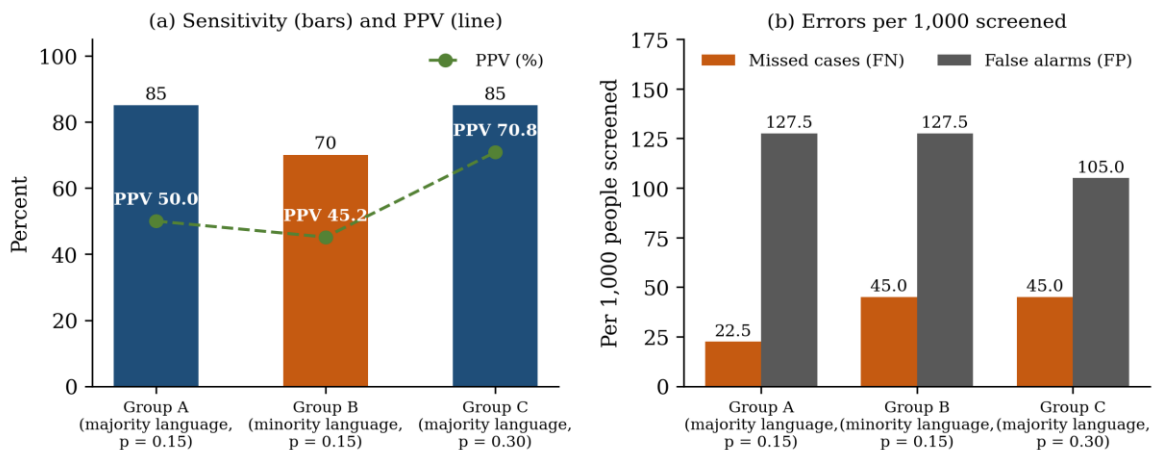


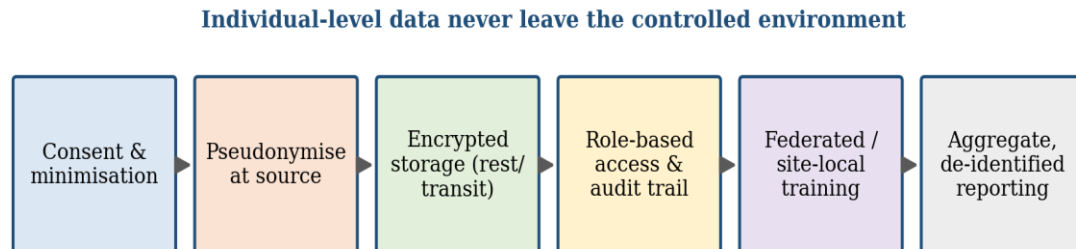
Fig. 3. Analytic fairness demonstration for a hypothetical model. (a) Sensitivity and positive predictive value by subgroup. (b) Missed cases and false alarms per 1,000 people screened, computed from equation (1) and standard formulas.

VI. PRIVACY-PRESERVING DATA PIPELINE

Fig. 4 outlines the proposed data pipeline. Consent and data minimisation come first, followed by pseudonymisation at the point of collection, encryption at rest and in transit, and role-based access with audit trails. Where data from several institutions are needed, federated learning allows models to be trained without moving individual records out of each site [24]. Outputs are released only as aggregate,



de-identified results. Raw audio and free text should be deleted after feature extraction wherever the approved protocol permits.



Raw audio and free text deleted after feature extraction where the protocol permits; retention periods defined in advance

Fig. 4. Privacy-preserving data pipeline for AI-assisted mental health research, in which individual-level data remain within a controlled environment.

VII. DISCUSSION

The framework reframes ethics as a set of verifiable engineering and clinical practices rather than a statement of values [6]. It also clarifies accountability. The treating professional remains responsible for diagnosis and care. Developers are responsible for documented performance, subgroup results and timely disclosure of limitations. Institutions are responsible for governance, monitoring and incident reporting [25]. Reporting under TRIPOD+AI [9] and staged clinical evaluation under DECIDE-AI [10] make these responsibilities visible to reviewers. In multilingual, resource-limited settings such as North-East India, the justice principle calls for special attention to language representation, digital access and culturally valid instruments, so that AI does not widen existing gaps in care [26].

1. Limitations

Risk ratings are qualitative and context-dependent. The fairness demonstration uses hypothetical parameters chosen to illustrate the mechanism, not observed results. The framework has not yet been tested in an ethics-committee review or a deployment study, which is the next planned step.

VIII. CONCLUSION

Responsible AI-assisted diagnosis of depression and anxiety depends on more than accuracy. It requires specific consent, safety protocols, calibrated and explained outputs, subgroup fairness, strong privacy protection and clear human accountability throughout the lifecycle. The proposed principle-to-practice framework, risk matrix and privacy pipeline give researchers and clinicians a concrete basis for building and reviewing mental health AI that people can trust.



Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

1. GBD 2019 Mental Disorders Collaborators, "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019," *Lancet Psychiatry*, vol. 9, no. 2, pp. 137–150, 2022, doi: 10.1016/S2215-0366(21)00395-3.
2. A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: A meta-analysis," *Lancet*, vol. 374, no. 9690, pp. 609–619, 2009, doi: 10.1016/S0140-6736(09)60879-5.
3. A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychol. Med.*, vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.
4. Y. Park, S. Park, and M. Lee, "Effectiveness of artificial intelligence in detecting and managing depressive disorders: Systematic review," *J. Affect. Disord.*, vol. 361, pp. 445–456, 2024, doi: 10.1016/j.jad.2024.06.035.
5. F. Zafar et al., "The role of artificial intelligence in identifying depression and anxiety: A comprehensive literature review," *Cureus*, vol. 16, no. 3, Art. no. e56472, 2024, doi: 10.7759/cureus.56472.
6. B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nat. Mach. Intell.*, vol. 1, no. 11, pp. 501–507, 2019, doi: 10.1038/s42256-019-0114-4.
7. D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care—Addressing ethical challenges," *N. Engl. J. Med.*, vol. 378, no. 11, pp. 981–983, 2018, doi: 10.1056/NEJMp1714229.
8. I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical machine learning in healthcare," *Annu. Rev. Biomed. Data Sci.*, vol. 4, pp. 123–144, 2021, doi: 10.1146/annurev-biodatasci-092820-114757.
9. G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, Art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.
10. B. Vasey et al., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *Nat. Med.*, vol. 28, no. 5, pp. 924–933, 2022, doi: 10.1038/s41591-022-01772-9.
11. W. N. Price II and I. G. Cohen, "Privacy in the age of medical big data," *Nat. Med.*, vol. 25, no. 1, pp. 37–43, 2019, doi: 10.1038/s41591-018-0272-7.
12. B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics," *BMC Med.*, vol. 17, Art. no. 230, 2019, doi: 10.1186/s12916-019-1466-7.
13. A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, 2006, doi: 10.1177/0272989X06295361.



14. A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Ann. Intern. Med.*, vol. 169, no. 12, pp. 866–872, 2018, doi: 10.7326/M18-1990.
15. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.
16. J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, Art. no. 310, 2020, doi: 10.1186/s12911-020-01332-6.
17. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
18. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 4765–4774, doi: 10.48550/arXiv.1705.07874.
19. M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.
20. M. P. Sendak, M. Gao, N. Brajer, and S. Balu, "Presenting machine learning model information to clinical end users with model facts labels," *npj Digit. Med.*, vol. 3, Art. no. 41, 2020, doi: 10.1038/s41746-020-0253-3.
21. S. G. Finlayson et al., "The clinician and dataset shift in artificial intelligence," *N. Engl. J. Med.*, vol. 385, no. 3, pp. 283–286, 2021, doi: 10.1056/NEJMc2104626.
22. K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: A systematic review of frequency, effect mediators, and mitigators," *J. Am. Med. Inform. Assoc.*, vol. 19, no. 1, pp. 121–127, 2012, doi: 10.1136/amiajnl-2011-000089.
23. S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. Int. Workshop Softw. Fairness (FairWare)*, 2018, pp. 1–7, doi: 10.1145/3194770.3194776.
24. N. Rieke et al., "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, Art. no. 119, 2020, doi: 10.1038/s41746-020-00323-1.
25. C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, Art. no. 195, 2019, doi: 10.1186/s12916-019-1426-2.
26. V. Patel et al., "The Lancet Commission on global mental health and sustainable development," *Lancet*, vol. 392, no. 10157, pp. 1553–1598, 2018, doi: 10.1016/S0140-6736(18)31612-X.