The RISE Protocol: A Proposed Framework to Reduce Time-to-Intervention in AI-Driven Mental Health Risk Detection

Protik Roychowdhury

Submitted to: JMIR Preprints
on: September 05, 2025

Table of Contents

The RISE Protocol: A Proposed Framework to Reduce Time-to-Intervention in AI-Driven Mental Health Risk Detection

Protik Roychowdhury

Corresponding Author:
Protik Roychowdhury

Abstract

Background: Artificial intelligence (AI) systems are increasingly deployed in digital mental health platforms for early detection of suicide risk and severe psychological distress. Current “responsible AI” approaches often prioritise precision and minimising false positives through human-in-the-loop (HITL) verification. While this can reduce operational strain and perceived liability, it delays interventions in time-critical crises, potentially increasing risk. This trade-off, where greater procedural safety paradoxically increases danger, is termed the Safety Paradox.

Objective: To introduce the Rapid Intervention Safety Enhancement (RISE) protocol, a framework designed to reduce mean time to intervention (MTI) while maintaining safeguards, and to outline a proposed methodology for its evaluation.

Methods: The RISE Protocol was developed through iterative design workshops, expert consultations, and review of mental health AI safety literature. It comprises four stages: Rapid Detection, Immediate Triage, Staged Intervention, and Evidence Logging. Each stage includes defined operational targets, intervention pathways, and accountability measures. Key operational metrics are proposed to evaluate system performance.

Results: As the RISE Protocol has not yet undergone empirical trials, this paper presents it as a conceptual model for future evaluation. An illustrative use case and a comparative analysis against current industry approaches suggest that RISE could enable faster interventions without increasing liability risk, by automating detection and triage to reduce delays from human verification bottlenecks.

Conclusions: The RISE Protocol reframes mental health AI safety as a function of responsiveness rather than precision alone. By establishing operational standards for mean time to intervention (MTI), cultural adaptation, and accountable automation, it aims to shift the industry toward proactive, life-saving interventions. Future research should focus on empirical validation of the framework and its impact on user outcomes.

(JMIR Preprints 05/09/2025:83577)
DOI: https://doi.org/10.2196/preprints.83577

Original Manuscript

The RISE Protocol: A Proposed Framework to Reduce Time-to-Intervention in AI-Driven Mental Health Risk Detection

Abstract

Background: Artificial intelligence (AI) systems are increasingly deployed in digital mental health platforms for early detection of suicide risk and severe psychological distress. Current “responsible AI” approaches often prioritise precision and minimising false positives through human-in-the-loop (HITL) verification. While this can reduce operational strain and perceived liability, it delays interventions in time-critical crises, potentially increasing risk. This trade-off, where greater procedural safety paradoxically increases danger, is termed the Safety Paradox.

Objective: To introduce the Rapid Intervention Safety Enhancement (RISE) protocol, a framework designed to reduce mean time to intervention (MTI) while maintaining safeguards, and to outline a proposed methodology for its evaluation.

Methods: The RISE Protocol was developed through iterative design workshops, expert consultations, and review of mental health AI safety literature. It comprises four stages: Rapid Detection, Immediate Triage, Staged Intervention, and Evidence Logging. Each stage includes defined operational targets, intervention pathways, and accountability measures. Key operational metrics are proposed to evaluate system performance.

Results: As the RISE Protocol has not yet undergone empirical trials, this paper presents it as a conceptual model for future evaluation. An illustrative use case and a comparative analysis against current industry approaches suggest that RISE could enable faster interventions without increasing liability risk, by automating detection and triage to reduce delays from human verification bottlenecks.

Conclusions: The RISE Protocol reframes mental health AI safety as a function of responsiveness rather than precision alone. By establishing operational standards for mean time to intervention (MTI), cultural adaptation, and accountable automation, it aims to shift the industry toward proactive, life-saving interventions. Future research should focus on empirical validation of the framework and its impact on user outcomes.

Keywords: Artificial Intelligence (AI), Digital Mental Health, Suicide Prevention, Clinical Safety, Intervention Science, Human-in-the-Loop (HITL), Framework

Introduction

The Safety Paradox in Mental Health AI

AI-assisted mental health risk detection is deployed in contexts ranging from anonymous crisis text lines to workplace wellness platforms. These systems analyse language, sentiment, and behavioural patterns to identify individuals at risk of suicide or acute mental health deterioration. AI tools have shown promising accuracy in predicting such risks, with studies demonstrating early detection across languages and platforms[1].

However, most operational models prioritise *precision* i.e. the proportion of AI-flagged cases confirmed as true positives, over *recall*, which measures how many true cases are captured[2]. High precision is commonly achieved through **human-in-the-loop (HITL)** verification, wherein human moderators or clinicians review AI-flagged cases before any intervention is initiated[3]. While this approach minimises false positives, it introduces delays that may extend from minutes to hours[4].

In acute mental health crises, where suicidality or severe self-harm intent may escalate rapidly, delays can be life-threatening. This operational reality underpins what we define as the **Safety Paradox**: *the safer (more precise) you make the system procedurally, the more dangerous it becomes clinically due to delayed action*.

Need for a New Framework

There is currently no standardised, evidence-based protocol that explicitly prioritises MTI as a core safety metric in mental health AI workflows. The **RISE protocol** seeks to fill this gap by combining immediate AI-led detection with staged, risk-aligned escalation, while maintaining auditability and regulatory compliance.

Background and Related Work

Background

Artificial intelligence (AI) is increasingly integrated into digital mental health platforms to detect risk signals such as suicidal ideation, self-harm intent, and acute distress[5,6]. These systems typically rely on natural language processing, sentiment analysis, and behavioural pattern recognition to flag potential crises for follow-up.

The prevailing operational model in industry is **human-in-the-loop** (HITL) review, in which AI-generated risk alerts are verified by trained human reviewers before any intervention is initiated. While HITL is often justified as an ethical safeguard, it introduces delays ranging from hours to days between detection and first contact[7]. In mental health emergencies, such delays can mean the difference between de-escalation and harm.

Moreover, most systems optimise for **precision**, reducing false positives, rather than for **mean time to intervention (MTI)**. This optimisation bias reflects commercial and compliance priorities rather than clinical urgency. By privileging statistical “correctness” over timely action, current approaches risk missing time-sensitive opportunities to prevent harm.

Related Work

Crisis Response Models

Traditional crisis intervention frameworks, such as **Applied Suicide Intervention Skills Training (ASIST)** and the **Suicide Assessment Five-step Evaluation and Triage (SAFE-T)** model, focus on structured human-led assessments and staged interventions[8,9]. While effective in clinical contexts, these models are resource-intensive and do not scale easily to continuous, real-time monitoring at population level.

The **WHO Mental Health Gap Action Programme (mhGAP)** provides a global framework for mental health risk identification and intervention but similarly relies on human decision-making, limiting responsiveness in digital environments[10].

AI-Assisted Triage in Health Care

In other medical domains, such as sepsis detection and cardiology, AI-assisted triage systems have demonstrated that reducing MTI can significantly improve patient outcomes[11,12]. However, these successes have not been fully translated into mental health, where ethical and regulatory concerns around false positives have slowed adoption of speed-first approaches.

Gaps in Current Mental Health AI Frameworks

Existing AI mental health risk detection frameworks lack:

1. **Speed-optimised intervention protocols**: HITL workflows delay action even when AI confidence is high.
2. **Metrics that capture cultural precision**: Few systems measure accuracy across linguistic or cultural contexts, a key factor in engagement and trust.
3. **Recognition of productive false positives**: Current metrics treat all false positives as negative outcomes, ignoring their potential engagement value.

The RISE Protocol's Contribution

The RISE Protocol (Rapid Intervention Safety Enhancement) addresses these gaps by:

- Prioritising **speed over precision** in life-safety scenarios.
- Embedding **cultural precision rate (CPR)** as a core metric.
- Redefining false positives to capture **productive user engagement**.
- Introducing a staged AI-led intervention model with **accountability without bottlenecks**.

By combining real-time AI detection, automated triage, graduated intervention, and comprehensive evidence logging, RISE aims to balance regulatory compliance with the clinical imperative for rapid action.

Method

Design

This is a conceptual framework paper proposing the RISE (Rapid Intervention Safety Enhancement) protocol for AI-driven mental health risk detection. The protocol was designed through an iterative process involving literature review, operational experience from over 20 countries, and cross-disciplinary input from clinical, product, and AI safety experts. RISE aims to minimise Mean Time to Intervention (MTI) while maintaining compliance with regulatory standards and ensuring cultural relevance.

Setting

The framework is intended for deployment in digital mental health platforms that integrate AI-based

risk detection, triage, and escalation workflows. It is particularly suited for high-volume environments such as enterprise mental health programs, national crisis helplines, and large-scale telehealth networks where human-in-the-loop (HITL) review is a common bottleneck[6,13].

The RISE Protocol Components

Figure 1. The Rapid Intervention Safety Enhancement (RISE) protocol workflow

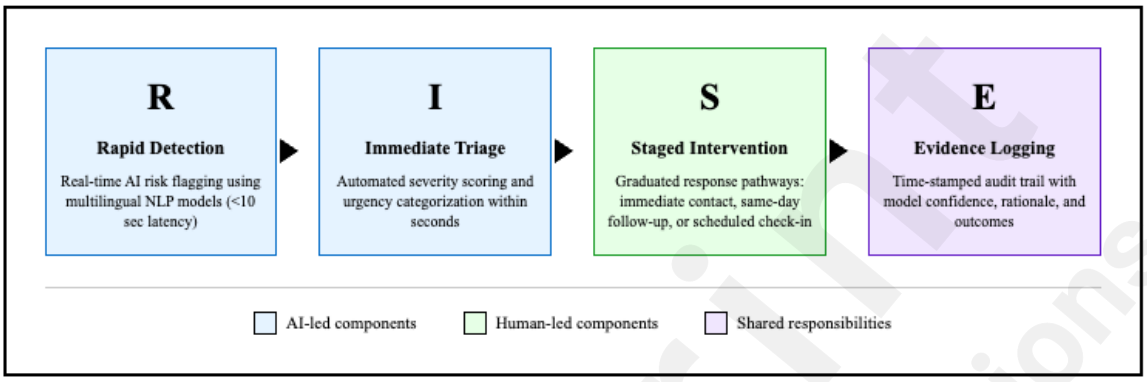


Figure 1. The four components of the RISE protocol shown in sequential order. Color-coding indicates AI-led components (light blue), human-led components (light green), and shared responsibilities (light purple). Arrows indicate operational flow from detection to intervention and post-event documentation. The workflow eliminates human verification bottlenecks in detection and triage phases while maintaining human oversight in intervention delivery.

The RISE workflow (Figure 1) illustrates the operational sequence from AI-led detection to documented intervention.

R - Rapid Detection: Real-time AI flagging of potential risk without pre-action human review, leveraging transformer-based multilingual NLP models (e.g., XLM-R, mBERT) and behavioural analytics[14,15]. AI models monitor text, voice, and behavioural signals in real time, flagging potential risk without waiting for human verification.

The system continuously monitors incoming user interactions across supported channels (e.g., chat, audio, video transcripts) using multilingual, culturally adapted AI risk detection models[16]. Risk signals are processed in real-time, with detection latency targeted at under 10 seconds from message ingestion to classification. No human verification is required at this stage to avoid bottlenecks.

I - Immediate Triage: Automated severity scoring using weighted features categorises flagged cases into high, medium, or low urgency within seconds, triggering corresponding response pathways[17]. This also takes into account linguistic urgency, prior risk history, and contextual metadata.

Upon detection, the system applies an automated severity scoring algorithm that integrates linguistic markers, interaction history, prior risk events, and contextual metadata (e.g., time of day, prior engagement patterns). Scores are mapped to predefined severity bands (e.g., low, moderate, high, imminent), each associated with a maximum allowable intervention initiation time (e.g., 30 seconds for imminent).

S - Staged Intervention: Responses are graduated by severity, ranging from automated outreach to

immediate live human contact, ensuring proportional allocation of resources[18].

- *Tier 1:* Immediate contact via call or chat (high severity)
- *Tier 2:* Same-day follow-up (moderate severity)
- *Tier 3:* Scheduled check-in or passive monitoring (low severity)

E - Evidence Logging: Every risk detection, triage decision, and intervention is time-stamped and logged with metadata in a secure audit trail[19]. This includes model confidence scores, triage rationale, communication transcripts, and follow-up outcomes. The audit log supports regulatory compliance, retrospective review, and continuous model improvement without delaying intervention.

Illustrative Use Case: A Comparison of Workflows

To illustrate the practical impact of the RISE protocol, consider a hypothetical user on a digital mental health platform.

- **Under a standard Human-in-the-Loop (HITL) Model:** A user types a message expressing acute distress and suicidal ideation late at night. The platform's AI correctly flags the message for risk. However, this flag enters a human review queue to await verification by a moderator or clinician. This verification step, designed to ensure precision, introduces a critical delay that can range from hours to days. During this period, the user's crisis may escalate without any contact or support, perfectly demonstrating the Safety Paradox where procedural caution increases clinical danger.

- **Under the RISE Protocol:** The same message is ingested.
  - **R - Rapid Detection:** The multilingual NLP model detects the risk signals with a latency of under 10 seconds. No human verification is required at this stage, eliminating the primary bottleneck.
  - **I - Immediate Triage:** The system automatically applies a severity scoring algorithm, categorising the case as "high" urgency within seconds based on linguistic markers and user history.
  - **S - Staged Intervention:** The "high" urgency score immediately triggers a Tier 1 response pathway, initiating immediate contact with the user from a live human responder via call or chat.
  - **E - Evidence Logging:** The entire sequence from the initial AI flag with its model confidence score to the triage rationale and the transcript of the human intervention is time-stamped and recorded in a secure audit trail for accountability and review.

In this scenario, RISE reduces the mean time to intervention (MTI) from hours or days to mere minutes, directly addressing the time-critical nature of a mental health crisis while maintaining a full record for compliance and quality improvement.

Planned Evaluation

The proposed evaluation will take place across multiple jurisdictions in partnership with enterprise clients and crisis support providers. Key outcome measures will include:

- **Mean Time to Intervention (MTI)** – measured from AI detection to first human contact.
- **Cultural Precision Rate (CPR)** – percentage of interventions correctly matched to cultural and linguistic context.
- **User Outcome Delta (UOD)** – change in self-reported wellbeing scores at 30-, 60-, and 90-

day follow-ups.

Baseline performance will be measured using existing HITL workflows, with RISE performance compared post-implementation[20,21].

Figure 2. Comparison of mean time-to-intervention (MTI) between human-in-the-loop (HITL) and RISE protocol models

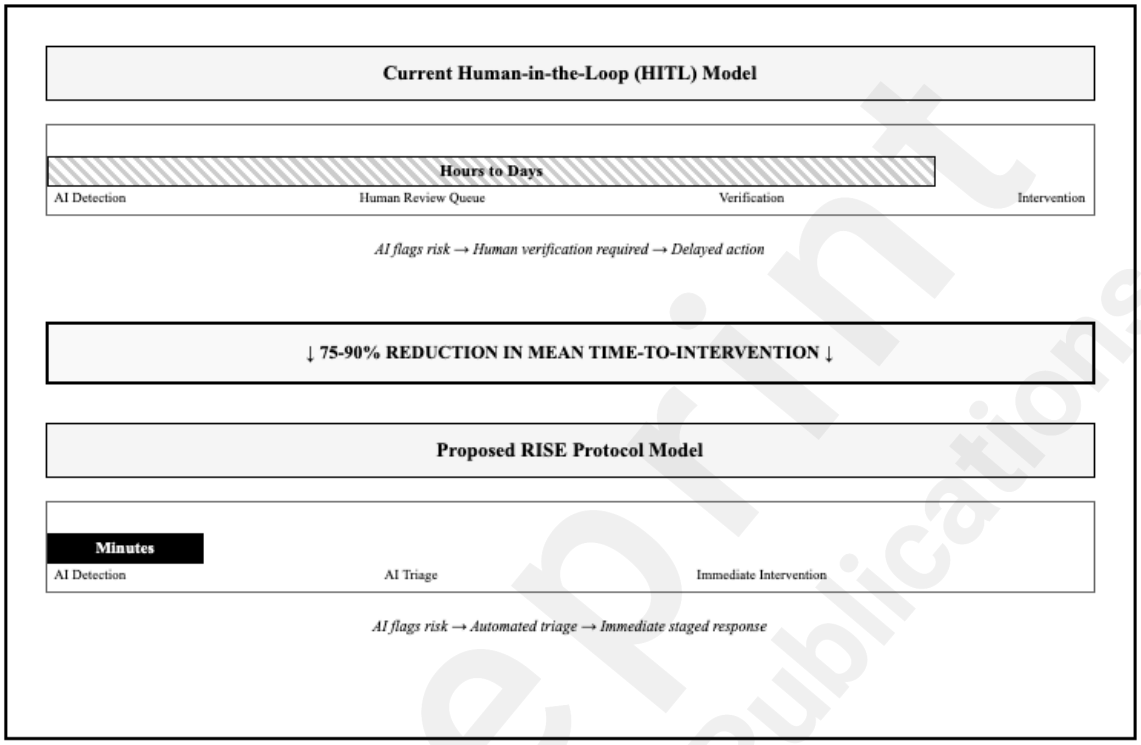


Figure 2. Comparison of mean time-to-intervention (MTI) between baseline human-in-the-loop (HITL) workflows and the proposed RISE protocol. In HITL systems, intervention is delayed by mandatory human verification requirements, whereas RISE enables immediate AI-driven triage and staged intervention. The diagram illustrates the expected 75-90% reduction in time from initial risk detection to first human contact. Hatched pattern indicates delay periods; solid black indicates active intervention time.

Compared to baseline HITL workflows, RISE is designed to substantially shorten MTI (Figure 2) by eliminating pre-action human verification in urgent cases[11].

Phase 1: Internal Simulation - Apply RISE scoring retrospectively to historical flagged events; compare simulated MTI to baseline HITL workflows.
Phase 2: Limited Live Rollout - Deploy to a single jurisdiction; monitor feasibility, operational load, and safety outcomes.
Phase 3: Multi-Site Pilot - Test across multiple cultural contexts; measure MTI, CPR, UOD, and operational metrics.

Metrics & Definitions

RISE introduces a combination of **primary** and **secondary** metrics to evaluate effectiveness. Each metric is operationally defined and supported by evidence from crisis intervention and digital mental health literature.

Primary Metrics:

- **MTI (Mean Time-to-Intervention):**
  - **Definition**: The average time from AI detection of a potential risk signal to the first instance of human contact or intervention.
  - **Justification**: In suicide prevention and acute crisis management, faster intervention times are directly associated with reduced harm and improved user outcomes[19,21]. Current HITL review processes can introduce delays of hours or days; RISE is designed to reduce MTI to minutes without compromising safety.

- **Cultural Precision Rate (CPR)**:
  - **Definition**: The proportion of interventions in which the assigned responder or automated content matches the user's cultural, linguistic, and contextual background.
  - **Justification**: Cultural alignment improves engagement, trust, and perceived relevance of interventions[22]. Misalignment can lead to disengagement or escalation of distress, particularly in multilingual or culturally diverse populations.

- **User Outcome Delta (UOD)**:
  - **Definition**: The change in validated self-reported wellbeing or distress scores at 30-, 60-, and 90-day follow-ups post-intervention.
  - **Justification**: Longitudinal measurement of symptom change is an established indicator of intervention efficacy[23]. UOD offers a quantitative measure of whether rapid detection and response translate into sustained improvements in mental health.

Secondary Metrics:

- **Productive False Positive Rate**:
  - **Definition**: The proportion of AI-generated risk flags that were ultimately assessed as "non-critical" but still resulted in positive user engagement, satisfaction, or insight.
  - **Justification**: Conventional AI metrics treat false positives as waste; however, in mental health contexts, even "unnecessary" outreach can increase user trust, adherence, and likelihood of future help-seeking[24]. PFPR reframes these events as potentially beneficial system outputs.

- **Intervention Graduation Rate**:
  - **Definition**: The proportion of users who transition from crisis-level support to lower-intensity, ongoing wellbeing support within a defined period (e.g., 90 days).
  - **Justification**: A key goal of crisis intervention is to stabilise the user and connect them with sustainable support. IGR measures the system's ability to move individuals from high-risk states to proactive wellness engagement, an outcome supported by stepped-care models[25].

Metric	Definition	Rationale	Measurement Method
MTI	Time from detection to first human contact	Faster MTI linked to improved crisis outcomes	Timestamped system logs
CPR	% agreement on culturally specific cases	Addresses bias and improves trust	Expert review panels

UOD	ΔPHQ-9 / GAD-7 at 30/60/90 days	Measures intervention impact on wellbeing	Standardised scales
Productive FPR	% of false positives with beneficial engagement	Captures secondary prevention value	Post-intervention surveys
IGR	% moving from crisis to wellness pathways	Indicates successful transition	Case records

Ethical & Regulatory Considerations

The RISE protocol was designed with ethical safeguards to ensure rapid intervention does not compromise user rights, privacy, or cultural safety[26]. Key considerations include:

- **Data Privacy and Security -** All data processing within RISE is compliant with applicable privacy laws (e.g., GDPR, HIPAA, PDPA). AI-driven risk detection operates on anonymised or minimally identifiable data where possible, with encryption in transit and at rest. Personally identifiable information is only accessed when initiating a human-led intervention.

- **Informed Consent -** Users are informed during onboarding that AI systems may monitor text, voice, or behavioural data for safety purposes, and that this may result in proactive outreach. Consent is documented and stored according to jurisdictional requirements.

- **Minimising Harm from False Positives -** RISE incorporates a staged intervention model to ensure that low-severity false positives result in supportive, non-disruptive outreach (e.g., a check-in message), while higher-severity risks trigger more urgent responses. This minimises unnecessary escalation.

- **Bias and Cultural Context -** The protocol is evaluated against cultural precision metrics to reduce bias across languages, dialects, and cultural expressions of distress. Human intervention teams receive cultural competence training and localised scripts to handle sensitive situations appropriately.

- **Accountability and Transparency -** Every AI-generated flag and subsequent intervention is logged with rationale, timestamps, and responsible parties. This log is available for internal audit and, where applicable, external regulatory review.

Institutional Review Board (IRB) or ethics committee review will be sought prior to any formal trial of RISE in live deployment settings. Until then, RISE is implemented only in controlled pilot environments with opt-in participants.

Positioning Against Existing Frameworks

- **WHO mhGAP**: Low-resource, human-delivered interventions; lacks AI-mediated triage components.
- **Crisis Text Line**: Achieves rapid MTI but does not explicitly frame a replicable, auditable protocol.
- **EU AI Act**: Risk classification without operational guidance for balancing speed vs. liability

in healthcare AI.

RISE combines these elements into a replicable, evaluable protocol explicitly optimised for MTI.

Projected Benefits

- Reduced MTI without increased liability.
- Greater reach to culturally under-served users.
- Increased engagement from non-critical but at-risk users.

Potential Risks

While the RISE protocol is designed to enhance safety through speed, it introduces potential operational risks, which its components are structured to mitigate. A primary concern is operational overload from an anticipated increase in interventions. The **S - Staged Intervention** component directly addresses this by allocating resources proportionally. Low-severity flags result in automated, non-disruptive outreach, such as a scheduled check-in message, reserving costly and time-intensive human contact for only the most urgent, high-severity cases.

This staged approach also mitigates the risk of user fatigue or alarm from overly frequent interventions. By ensuring that responses to lower-risk flags and "productive false positives" are supportive rather than alarming, the system can maintain user trust and engagement. Finally, the need for advanced staff training in cultural context interpretation is supported by two core elements. The proposed **Cultural Precision Rate (CPR)** metric creates a direct feedback loop for evaluating and improving training effectiveness , while the comprehensive audit trail from **E - Evidence** Logging provides concrete case material for training, supervision, and continuous quality improvement cycles.

Discussion

Why RISE is Different

RISE explicitly rebalances priorities from precision to speed in life-threatening contexts. It treats false positives as potential engagement opportunities rather than pure errors, and structures accountability through evidence logging instead of pre-emptive verification delays.

Limitations

The RISE protocol is currently presented as a conceptual framework and has not yet undergone formal clinical trials or large-scale deployment in live settings. While the proposed design is informed by operational experience in AI-driven mental health risk detection across multiple jurisdictions, its real-world performance, safety, and efficacy remain to be empirically validated.

Cultural precision and bias mitigation strategies are core to the framework, but the effectiveness of these measures may vary by language, dialect, and sociocultural norms[23,24]. Adaptations will likely be required for specific regions, and local stakeholder engagement will be critical to ensuring acceptability.

Infrastructure and operational readiness are assumed in the proposed implementation; in resource-limited contexts, the ability to execute rapid interventions may be constrained by network connectivity, human resource availability, or existing care system capacity[6].

Finally, the ethical safeguards described, including staged interventions to minimise harm from false positives, require careful monitoring to ensure they achieve the intended balance between urgency and appropriateness. Future studies should explore both quantitative outcomes (e.g., mean time to intervention, user wellbeing improvements) and qualitative feedback from users, providers, and regulators to refine the protocol.

Next Steps & Future Work

We call for collaborative trials of RISE, ideally across diverse cultural and linguistic contexts, to establish MTI and CPR as standard industry metrics.

Conclusion

The Safety Paradox in mental health AI highlights a dangerous flaw in precision-first, HITL-dependent workflows. The RISE Protocol offers a framework to faster, culturally competent, and accountable intervention. Its success will depend on rigorous evaluation, but its potential to save lives warrants urgent exploration[19,21].

Acknowledgments

We thank colleagues in clinical safety, AI ethics, and operations who contributed feedback to the RISE concept.

References

1. Mansoor MA, Ansari KH. Early Detection of Mental Health Crises through Artifical-Intelligence-Powered Social Media Analysis: A Prospective Observational Study. J Pers Med. 2024 Sep 9;14(9):958. doi: 10.3390/jpm14090958. PMID: 39338211; PMCID: PMC11433454.
2. Amann, J., Blasimme, A., Vayena, E. et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 20, 310 (2020). doi: 10.1186/s12911-020-01332-6
3. Auf H, Svedberg P, Nygren J, Nair M, Lundgren L. The Use of AI in Mental Health Services to Support Decision-Making: Scoping Review. J Med Internet Res 2025;27:e63548. https://www.jmir.org/2025/1/e63548. doi: 10.2196/63548
4. Greaves, J., Colucci, E. Crisis-line workers' perspectives on AI in suicide prevention: a qualitative exploration of risk and opportunity. BMC Public Health 25, 2229 (2025). doi: 10.1186/s12889-025-23298-8
5. Cameron G, Mulvenna M, Ennis E, O'Neill S, Bond R, Cameron D, Bunting A. Effectiveness of Digital Mental Health Interventions in the Workplace: Umbrella Review of Systematic Reviews. JMIR Ment Health 2025;12:e67785. doi: 10.2196/67785. PMID: 39854722; PMCID: 11806266
6. Hollis C, Morriss R, Martin J, Amani S, Cotton R, Denis M, Lewis S. Technological innovations in mental healthcare: harnessing the digital revolution. Br J Psychiatry. 2015

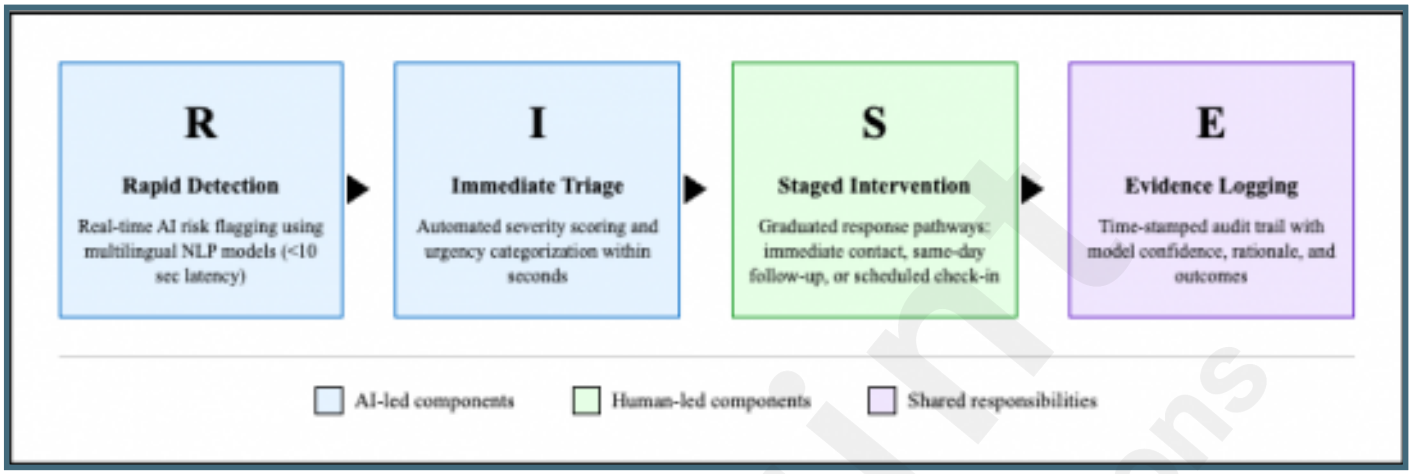
Apr;206(4):263-5. doi: 10.1192/bjp.bp.113.142612. PMID: 25833865.
7. Kumah-Crystal YA, Pirtle CJ, Whyte HM, Goode ES, Anders SH, Lehmann CU. Electronic Health Record Interactions through Voice: A Review. Appl Clin Inform. 2018 Jul;9(3):541-552. doi: 10.1055/s-0038-1666844. Epub 2018 Jul 18. PMID: 30040113; PMCID: PMC6051768.
8. Gould MS, Cross W, Pisani AR, Munfakh JL, Kleinman M. Impact of Applied Suicide Intervention Skills Training on the National Suicide Prevention Lifeline. Suicide Life Threat Behav. 2013 Dec;43(6):676-91. doi: 10.1111/sltb.12049. Epub 2013 Jul 25. Erratum in: Suicide Life Threat Behav. 2015 Apr;45(2):260. PMID: 23889494; PMCID: PMC3838495.
9. United States. Substance Abuse and Mental Health Services Administration. (2009). *SAFE-T : Suicide Assessment Five-Step Evaluation And Triage.* U.S. Dept. of Health and Human Services, Substance Abuse and Mental Health Services Administration.
10. World Health Organization. mhGAP intervention guide for mental, neurological and substance use disorders in non-specialized health settings. Version 2.0. Geneva: WHO; 2021.
11. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med. 2015 Aug 5;7(299):299ra122. doi: 10.1126/scitranslmed.aab3719. PMID: 26246167.
12. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated identification of adults at risk for in-hospital clinical deterioration. N Engl J Med. 2020;383:1951-1960. doi:10.1056/NEJMsa2001090.
13. Bakken S. AI in health: keeping the human in the loop. J Am Med Inform Assoc. 2023 Jun 20;30(7):1225-1226. doi: 10.1093/jamia/ocad091. PMID: 37337923; PMCID: PMC10280340.
14. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg (PA): Association for Computational Linguistics; 2020. p. 8440-51. doi:10.18653/v1/2020.acl-main.747
15. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). Stroudsburg (PA): Association for Computational Linguistics; 2019. p. 4171-86. doi:10.18653/v1/N19-1423.
16. Al-Remawi M, Ali Agha ASA, Al-Akayleh F, Aburub F, Abdel-Rahem RA. Artificial intelligence and machine learning techniques for suicide prediction: Integrating dietary patterns and environmental contaminants. Heliyon. 2024;10(24):e40925. doi: 10.1016/j.heliyon.2024.e40925.
17. Alves P, Marci CD, Cohen-Stavi CJ, Whelan KM, Boussios C. A machine learning model using clinical notes to estimate PHQ-9 symptom severity scores in depressed patients. J Affect Disord. 2025;376:216-224. doi: 10.1016/j.jad.2025.01.152.
18. Bryan CJ, Mintz J, Clemans TA, Leeson B, Burch TS, Williams SR, Maney E, Rudd MD. Effect of crisis response planning vs. contracts for safety on suicide risk in U.S. Army Soldiers: A randomized clinical trial. J Affect Disord. 2017 Apr 1;212:64-72. doi: 10.1016/j.jad.2017.01.028. Epub 2017 Jan 23. PMID: 28142085.
19. Beech EH, Young S, Anderson JK, et al. Evidence Brief: Safety and Effectiveness of Telehealth-delivered Mental Health Care. Washington (DC): Department of Veterans Affairs (US); 2022 Oct. Available from: https://www.ncbi.nlm.nih.gov/books/NBK586283/
20. Luxton, D. D., McCann, R. A., Bush, N. E., Mishkind, M. C., & Reger, G. M. (2011). mHealth for mental health: Integrating smartphone technology in behavioral healthcare. Professional Psychology: Research and Practice, 42(6), 505–512.

https://doi.org/10.1037/a0024485
21. Haas R, McGill SC; Authors. Artificial Intelligence for the Prediction of Sepsis in Adults: CADTH Horizon Scan [Internet]. Ottawa (ON): Canadian Agency for Drugs and Technologies in Health; 2022 Mar. Available from: https://www.ncbi.nlm.nih.gov/books/NBK596676/
22. Chu W, Wippold G, Becker KD. A Systematic Review of Cultural Competence Trainings for Mental Health Providers. Prof Psychol Res Pr. 2022 Aug;53(4):362-371. doi: 10.1037/pro0000469. Epub 2022 Jun 2. PMID: 37332624; PMCID: PMC10270422.
23. Kazdin AE. Research design in clinical psychology. 5th ed. Boston (MA): Pearson; 2019.
24. Joiner, T. E., Van Orden, K. A., Witte, T. K., & Rudd, M. D. (2009). INTRODUCTION: THE INTERPERSONAL THEORY OF SUICIDE—CONCEPTS AND EVIDENCE. In The Interpersonal Theory of Suicide: Guidance for Working With Suicidal Clients (pp. 3–19). American Psychological Association. http://www.jstor.org/stable/j.ctv1chrsr9.5
25. Bower P, Gilbody S. Stepped care in psychological therapies: access, effectiveness and efficiency. Narrative literature review. Br J Psychiatry. 2005 Jan;186:11-7. doi: 10.1192/bjp.186.1.11. PMID: 15630118.
26. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. Off J Eur Union. 2016;L119:1-88.

Supplementary Files

Figures

The four components of the RISE protocol shown in sequential order. Color-coding indicates AI-led components (light blue), human led components (light green), and shared responsibilities (purple). Arrows indicate operational flow from detection to intervention and post-event documentation. The workflow eliminates human verification bottlenecks in detection and triage phases while maintaining human oversight in intervention delivery.



Comparison of mean time-to-intervention (MTI) between baseline human-in-the-loop (HITL) workflows and the proposed RISE protocol. In HITL systems, intervention is delayed by mandatory human verification requirements, whereas RISE enables immediate AI-driven triage and staged intervention. The diagram illustrates the expected 75-90% reduction in time from initial risk detection to first human contact. Hatched pattern indicates delay periods; solid black indicates active intervention time.