Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Human-Centered AI in Crisis Counseling: A Three-Tier Governance Framework

Ambika Thangaraju

ambikathangaraju93@gmail.com

All India Institute of Medical Sciences

Method Article

Keywords: Crisis counseling, suicide prevention, artificial intelligence, human-in-the-loop AI, clinical governance, digital mental health

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11148168/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Crisis counselors must rapidly manage critical, complex demands under high pressure: conducting suicide risk assessments, providing emotional support, managing paperwork, and handling caseloads. Artificial intelligence (AI) has the potential to assist counselors with many aspects of their jobs, but ethical and clinical concerns surround its integration. AI has been studied for suicide-risk prediction, but this evidence does not by itself establish how AI should be governed within live crisis-counseling workflows (Greaves and Colucci, 2025). Established models for crisis counseling were distilled into a framework comprising three tiers, from pre-connection to post-call follow-up, incorporating the ABC Crisis Intervention Model, the Collaborative Assessment and Management of Suicidality, Columbia-Suicide Severity Rating Scale-based assessment, the Stanley-Brown Safety Planning Intervention, Motivational Interviewing, and core principles of Psychological First Aid. Tier 1 positions human-centered AI within the clinical workflow as a consultative support tool rather than an autonomous decision-maker, providing administrative, documentation, information-retrieval, and workflow support under human oversight. It does not assign suicide risk levels, determine clinical disposition, initiate emergency interventions, diagnose, or override clinical decisions. Tier 2 keeps clinical judgment, suicide risk assessment, and clinical decision-making under human clinical authority, using established evidence-based crisis-counseling processes. Tier 3 establishes organizational and clinical oversight, supervision, and quality-improvement measures to protect the mental health of crisis-line staff and volunteers. By separating clinical judgment from AI functions while maintaining human-in-the-loop oversight, this framework embeds governance protocols across the full crisis-care workflow without shifting decision-making authority away from the crisis clinician.

Highlights

- A three-tier governance framework for AI in crisis counseling is proposed.
- Tier 1 AI supports workflow, documentation, and administration under human oversight.
- Tier 2 keeps suicide risk assessment and disposition under human authority.
- A Clinical Uncertainty Pathway and Reality-Testing Override are introduced.
- Emergency escalation triggers AI lockout with audit logging.

1. Introduction

In crisis counseling, professionals must make rapid decisions under tight time constraints and limited information. A missed risk indicator can cause serious harm, while poorly targeted responses may consume limited service capacity and risk disengagement.

Crisis services, which have evolved with changing service models and demands (Roberts, 2005), can reduce mental distress and suicidal distress (Mishara et al., 2016; Hoffberg et al., 2020), but crisis-line

work also places substantial demands on staff and may be affected by workload, emotional strain, and variability in practice (Mishara and Daigle, 1997; Gould et al., 2007). AI has increasingly been studied for suicide-risk prediction and analysis of language or behavioral signals (Belsher et al., 2019; Coppersmith et al., 2018; Bernert et al., 2020), but these approaches do not establish how AI should be governed within crisis-counseling workflows (Greaves and Colucci, 2025).

We conducted a focused conceptual synthesis of established crisis-intervention and suicide-risk-management models, along with relevant literature on AI-supported mental-health care and clinical governance. Candidate models (including ASIST and Mental Health First Aid) were screened against three criteria — established empirical or clinical use, relevance to crisis-line or acute crisis care, and coverage of complementary stages of crisis intervention — and excluded where any criterion was unmet. The selected models were then mapped onto a proposed crisis-counseling workflow, with AI functions assigned based on their potential to support, rather than determine, clinical decisions. This approach follows the conceptual-framework methodology described by Jabareen (2009) and is intended to generate a testable governance framework rather than a systematic review.

Designed for live phone and text crisis services, this framework supports rather than replaces clinical training, supervision, and professional judgment. The six models and their placement within the clinical workflow are summarized in Table 1

Table 1
Overview of the six clinical models and assessment/intervention approaches synthesized into the human-centered AI (HCAI) framework, with each model's primary purpose and its placement within the Tier 2 clinical workflow.

Model	Primary purpose	Framework application
ABC Crisis Intervention Model	Three-phase intervention (Achieving contact, Boiling down the problem, Coping)	Step 0 (emotional regulation), Step 4A (low-risk intervention)
Collaborative Assessment and Management of Suicidality (CAMS)	Collaborative assessment of suicidal drivers, suicidal intent, planning, access to means, and reasons for living, with formulation and treatment planning	Step 3 (core risk assessment)
Columbia-Suicide Severity Rating Scale (C-SSRS)	Standardized assessment of suicidal ideation and behavior	Step 1.1 (structured risk assessment: ideation, intent, plan, means), Step 2 (severity characterization of ideation and behavior)
Stanley-Brown Safety Planning Intervention	Collaborative development of coping strategies and emergency contacts	Step 4B (moderate-risk intervention)
Motivational Interviewing	Person-centered communication to strengthen motivation for change	Step 4A (low-risk intervention)
Psychological First Aid (WHO et al., 2011)	Immediate safety prioritization and stabilization	Step 0 (safety acknowledgment)

2. A three-tier governance architecture for AI-supported crisis counseling

- Tier 1 (AI Workflow Support): administrative tasks before, during, and after calls, without making, determining, or autonomously executing clinical decisions.
- Tier 2 (Human Clinical Decision-Making): human counselors retain primary clinical authority, with escalation to supervisors or other designated clinical professionals when required.
- Tier 3 (Organizational Governance): supervision, quality assurance, and safeguards for counselor well-being.

The governing distinction is between assistance and determination: Tier 1 may surface, retrieve, organize, document, or prompt information, but it may not independently interpret that information as a clinical determination, initiate a clinical intervention, or alter the caller's clinical pathway.

For this framework, 'human clinical authority' refers to the appropriately trained and authorized professional or designated crisis-service staff member responsible for the relevant clinical or safeguarding decision; the specific role and escalation pathway will vary by service. Figure 1 outlines the complete architecture.

3. Tier 1: The AI workflow support tier

Tier 1 is intended to reduce administrative and cognitive burdens, support consistent protocol adherence, strengthen documentation, and maintain continuity of care.

3.1. Pre-connection support

To ease hold times, optional pre-connection support may acknowledge the caller in their preferred language, offer brief pre-recorded reassurance or resource information, and transfer the call once a counselor connects. Because synthetic audio may be experienced as impersonal or inappropriate by some distressed callers, services should evaluate its use carefully before adopting it in place of pre-recorded content.

Callers must be told immediately they are interacting with an automated system, offered silence as an alternative, and prioritized for human connection according to predefined, service-approved criteria rather than any AI interpretation of clinical state. Where feasible, services could adopt a one- to two-minute maximum transfer target, with a contingency procedure when staffing prevents this.

3.2. In-interaction support functions

Functions include flagging predefined linguistic or vocal features that may warrant counselor attention, identifying incomplete documentation, retrieving resources, drafting notes for review, and providing translation or multilingual support where validated for the relevant language and context.

Because this tier may engage with caller language or vocalization, callers must be told of its use at the start of the interaction, in a brief scripted statement, with fuller detail available in the service's published protocol. This notification is separate from follow-up consent, and callers should be offered an immediate opt-out from nonessential language or vocal-feature analysis; opting out must stop that analysis and be logged independently.

3.3. Post-interaction follow-up

The system schedules follow-up according to the counselor's documented risk assessment and a predefined, counselor-approved protocol, delivered through the most available consented channel.

Because decision-making capacity fluctuates during crisis, consent cannot be static: callers should consent to the specific automated follow-up channel and frequency before any automated contact, and may withdraw that consent at any time. Opting out stops all automated contact immediately, though secure access to historical records is maintained. Non-response should never be equated with voluntary withdrawal; where clinically indicated, it may instead trigger a predefined follow-up or escalation protocol established by the service.

3.4. Operational constraints on AI workflow support

All AI-generated content must be labeled and reviewed by a counselor before entering the clinical record, and counselors bear sole responsibility for evaluating discrepancies between AI-generated content and their own clinical assessment.

Tier 1 limits also guard against two risks: automation complacency, a growing reliance on system outputs (Parasuraman and Riley, 1997; Parasuraman and Manzey, 2010), and fabrication, in which AI-generated documentation reads fluently while containing invented information. The boundaries in Table 2 and regular audits (Section 5.4) address both risks.

Table 2

Operational constraints defining the boundary between assistive and prohibited AI functions within the HCAI framework. The constraints are intended to keep AI a workflow support tool while clinical authority, accountability, and emergency decision-making remain exclusively under human control.

AI may (supportive functions)	AI may not (prohibited functions)
Organize and surface relevant clinical information	Diagnose mental health conditions
Retrieve organization-approved evidence-based protocols and checklists	Assign suicide risk level or category
Flag predefined linguistic or vocal features that may warrant counselor attention	Determine clinical disposition or pathway selection
Draft clinical documentation for mandatory counselor review	Initiate emergency interventions or contact emergency services
Schedule follow-up contacts according to counselor-determined risk and a predefined, counselor-approved protocol	Override or contradict counselor clinical judgment
Highlight incomplete or missing documentation	Make autonomous clinical decisions of any kind
Support workflow consistency through structured prompts	Require clear identification of AI-generated content in clinical records

3.5. Technical continuity, capability boundaries, and linguistic equity

The framework does not mandate any specific AI model or platform; the organization, not the technology, remains responsible for ensuring compliance with Table 2 and maintaining human clinical control.

Expressions of distress vary across cultures and languages; callers often rely on metaphors, and automated analysis is less reliable in underrepresented languages, with dialects and code-switching further reducing transcription performance (Joshi et al., 2020). Bias and equity failures in deployed clinical algorithms are well documented (Char et al., 2018; Obermeyer et al., 2019). Organizations should establish evidence of acceptable performance for each language and dialect in use and monitor performance after implementation, with human review of translated or interpreted content. Tier 1 functions should be introduced in phases according to their potential clinical impact, beginning with low-consequence documentation and administrative tasks before deploying functions that interpret caller language, generate clinical prompts, or influence follow-up.

4. Tier 2: Human clinical decision-making

Tier 2 provides a sequential clinical workflow guiding each crisis encounter from initial contact to resolution (Fig. 2). Each step defines a core clinical decision made by the counselor, supported by Tier 1 activities; Table 3 sequences when each AI function activates, while Table 2 specifies what AI may and

may not do at any point in that sequence. Clinical assessment and intervention remain strictly under human control.

4.1. Steps 0–3: Safety, structured assessment, and core risks

Following the ABC Crisis Intervention Model (Kanel, 2019) and the safety-prioritization and stabilization principles of Psychological First Aid (World Health Organization et al., 2011), safety and emotional regulation are established simultaneously, since acute distress impairs a person's ability to assess risk independently (Step 0).

Structured assessment (Step 1) has three parts: standard risk assessment covering ideation, intent, plan, and means (Step 1.1); a medical red-flag gate (Step 1.2); and a substance/state reliability gate for intoxication or withdrawal (Step 1.3). A positive medical red-flag gate should trigger the service's established emergency medical evaluation or emergency-services pathway, according to local protocol. These gates are intended to identify potentially reversible medical conditions and altered states; complications at either step lead to the Clinical Uncertainty Pathway (Section 4.2) or supervisor and governance review (Section 5.1).

Step 1.1 establishes whether ideation, intent, plan, and means are present; Step 2 then applies the C-SSRS to characterize the severity and intensity of any ideation and behavior identified, which Step 1.1 does not grade (Posner et al., 2011).

4.2. Uncertainty and reality-testing safeguards

Crisis calls rarely present complete information: missing detail, unclear accounts, and ambivalence are endemic to crisis care, as they are to diagnostic reasoning generally (Croskerry, 2009). For this framework, clinical uncertainty refers to situations in which key information is missing, conflicting, unreliable, or insufficient to support a confident risk formulation or disposition, warranting deferral rather than emergency response. Rather than impose an unsubstantiated risk classification, this framework defers to supervisory review, which may revise or supersede the counselor's assessment. Where no supervisor is available, the procedures in Section 5.1 apply.

Loss of reality testing calls for a different approach; acute disturbance of reality testing may make remote risk assessment and routine low-intensity pathways unreliable. For this framework, the Reality-Testing Override criterion is met when the counselor observes disorganized speech, apparent hallucinatory or delusional content, or loss of orientation to the conversation sufficient to prevent shared understanding of the caller's immediate circumstances. When the criterion is met, routine low-intensity pathways are bypassed in favor of prompt clinical assessment and crisis management. The disturbance is evaluated by an authorized clinical or supervisory professional, whose declaration of an emergency activates the AI lockout protocol under Section 5.2.

4.3. Steps 4–7: Intervention, review, and closure

Following risk assessment, the counselor selects one of three pathways: Step 4A, low-risk, uses the ABC Model and motivational interviewing (Miller and Rollnick, 2013). Step 4B, moderate-risk, employs the Stanley-Brown Safety Planning Intervention (Stanley and Brown, 2012), addressing risk factors, means safety, and reassessment before the interaction ends; safety-planning-type interventions have shown reduced suicidal behavior relative to usual care (Stanley et al., 2018; Nuij et al., 2021). Step 4C, high-risk, applies when the counselor determines that acute suicide risk warrants escalation. The counselor should involve the supervisor and follow the service's emergency and safeguarding procedures (American Psychiatric Association, 2003), including any legal or duty-to-protect requirements applicable in the service's jurisdiction; the supervisor or other authorized clinical professional then declares an emergency, which activates the AI lockout protocol (Section 5.2).

Any interaction triggering shared responsibility (Step 4C, clinical uncertainty, or an unresolved gate) proceeds to supervisor and governance review (Step 5). After an emergency escalation, counselors receive a decompression period during which the supervisor checks in (Step 6); Tier 1's role is limited to summarizing the interaction, not reviewing performance.

Where no emergency action has been taken, Step 7 closes the encounter by summarizing the key concerns discussed, agreed coping or support strategies where appropriate, and next steps. Where emergency action has been taken, the encounter closes under Step 5, and documentation and handover are completed there.

Table 3

Tier 2 clinical workflow steps and their corresponding Tier 1 AI workflow support functions. Rows marked cross-cutting are safeguards that may activate at any point rather than sequential steps.

Tier 2 clinical step	Tier 1 AI workflow support
Step 0: Safety acknowledgment and emotional regulation	Language identification; where deployed under the phased-introduction rule in Section 3.5, logging of predefined voice or language features that may warrant counselor attention
Step 1.1: Structured suicide risk assessment (C-SSRS-informed)	C-SSRS protocol retrieval; structured assessment template
Step 1.2: Medical red-flag gate	Red-flag checklist retrieval and prompt; supervisor notification according to counselor- or protocol-defined criteria
Step 1.3: Substance/state reliability gate	Intoxication/withdrawal checklist prompt; documentation of uncertainty
Step 2: C-SSRS severity characterization of ideation and behavior	Structured response capture; assessment completeness check
Step 3: Core risk assessment (CAMS-aligned)	Domain-specific prompts (suicidal drivers, intent, planning, means, and reasons for living)
Clinical Uncertainty Pathway (cross-cutting)	Documentation of counselor-identified uncertainty; reassessment reminder; supervisor notification according to protocol; AI lockout does not apply
Reality-Testing Override (cross-cutting)	Emergency protocol retrieval; AI lockout on emergency declaration by an authorized clinical or supervisory professional (Section 5.2)
Step 4A: Low-risk pathway	Local resource navigation; retrieval of organization-approved coping resources; counselor-selected coping strategies
Step 4B: Moderate-risk pathway	Safety plan template; retrieval of organization-approved safety-planning and means-safety resources for counselor review; follow-up scheduling
Step 4C: High-risk pathway	AI lockout on emergency declaration by an authorized clinical or supervisory professional
Step 5: Supervisor and governance review	Documentation of clinical rationale; audit-ready record maintenance
Step 6: Counselor decompression	Neutral interaction summary; workload monitoring
Step 7: Post-call closure and documentation (non-emergency encounters)	Draft clinical note generation; missing documentation alert; agreed next-steps summary
Post-interaction follow-up (not a numbered step)	Follow-up scheduling based on counselor-approved protocol; non-response flagged to the counselor according to predefined escalation criteria; counselor-approved closing message

5. Tier 3: Organizational governance

5.1. Supervision and shared responsibility

Structured counselor–supervisor consultation is an important safeguard for high-risk or uncertain cases. Within this framework, supervisor review (Step 5 of the Tier 2 sequence, governed by the Tier 3 policies below) is required for Step 4C, clinical uncertainty, and any gate not resolved as a clear negative. Tier 1 supports this by documenting clinical logic and maintaining an audit trail, while human supervision remains a separate, parallel pathway; the supervisor's title varies by service, but the governance role is constant.

Services should have a documented backup process for reaching an alternate supervisor or physician, with predefined response-time targets calibrated to local service capacity and governance requirements. A prolonged inability to reach an on-call supervisor should be treated as a staffing-governance risk requiring remedy, not as a permitted deviation from the review requirement.

To prevent review backlogs, organizations should set tiered review priorities; within the proposed framework, Step 4C, the Reality-Testing Override, and emergency actions would warrant immediate review, whereas clinical uncertainty and unresolved gates could be reviewed within the same shift. Supervisor workload and review time should be measured as part of Tier 3 auditing.

5.2. Emergency escalation and AI lockout

AI lockout is a governance control rather than a clinical intervention. An emergency declaration may be made only by an authorized clinical or supervisory professional, and is required at Step 4C and when the Reality-Testing Override criterion is met; the Clinical Uncertainty Pathway does not activate lockout. On declaration, Tier 1 enters a lockout state that prevents any new AI-generated analysis or output. Lockout should remain in force until the emergency is resolved and the authorized clinical or supervisory professional completes the relevant handoff; if activation cannot be technically confirmed, AI output must remain disabled and human-only operation must continue. Pre-lockout outputs remain labeled and available for record review but carry no additional authority. Only explicitly authorized nonclinical record-keeping functions may continue during lockout, and their initiation is logged in the standard audit trail.

5.3. Counselor decompression and well-being

Workflows should include counselor decompression and supervisor check-in after emergency escalation (Step 6), and after other high-risk interactions where the service provides for this, with Tier 1 in neutral-summary mode. Evidence on compulsory debriefing is mixed to negative (Rose et al., 2002), but organizations should ensure access to protected recovery time or peer support. Compassion fatigue and secondary traumatic stress are well-documented hazards for crisis counselors (Figley, 2002; Bride et al., 2007), tracked as part of Tier 3 auditing.

5.4. Audit, regulatory positioning, and accountability

Routine audits should review every interaction escalated to Step 5, that is, Step 4C, the Reality-Testing Override, clinical uncertainty, and unresolved gates, plus a sample of non-escalated interactions, with AI model updates revalidated rather than assumed unchanged.

Where a Tier 1 function meets the definition of a medical device, documentation of intended use, regulatory classification, and evidence of ongoing compliance are required. Classification depends on intended use, degree of clinical influence, and jurisdiction; organizations should determine it before deployment and after substantive system changes.

Ultimate clinical responsibility remains with human counselors and supervisors. Adverse-event reviews must examine AI's role without attributing automated culpability, and organizations must maintain incident logs distinguishing human error from system failure.

Table 4
Key practical barriers to implementing human-centered artificial intelligence in crisis counseling services, with accompanying mitigation strategies.

Barrier	Mitigation strategy
Staff resistance or anxiety about AI replacing clinical judgment	Emphasize the governance structure (AI lockout, Table 2 constraints); involve counselors in pilot design; provide transparent communication about AI's limited role
IT infrastructure limitations (legacy systems, integration costs)	Phased implementation starting with documentation support and post-call drafting (lowest integration burden); consider cloud-based or vendor-managed solutions only where security, privacy, data governance, interoperability, and regulatory requirements are demonstrably satisfied
Data privacy and security concerns	Conduct data protection impact assessments; implement role-based access controls; ensure encryption and audit logging; comply with applicable privacy regulations
Cost constraints for smaller crisis services	Prioritize tools with appropriate evidence, security, governance support, and affordability; consider collaborative procurement across service networks; begin with lower-risk Tier 1 functions such as documentation drafting
Staff training burden	Embed AI training into existing clinical supervision and continuing education; create 'AI champion' roles among staff for peer support; develop concise quick-reference guides

6. Discussion

6.1. Relation to prior work and strengths

A substantial body of prior work has focused on algorithmic prediction or the safety of AI-generated responses; this framework instead focuses on how AI-supported functions can be governed within the live crisis-counseling workflow while preserving human clinical authority. Explicit governance architectures for AI in psychological counseling are rare; the closest comparator (Erdemir and Sumbas,

2026) addresses tiered oversight for suicide-related communication without detailing workflow integration.

Related work instead evaluates AI-response safety in suicide-related dialogue (Wei et al., 2026), addressing output safety rather than organizational architecture.

The framework's key strength relative to risk-prediction work is its attention to accountability: it assigns human responsibility for reviewing and documenting discrepancies between AI-generated content and the counselor's assessment. Its principal proposed contributions are the Clinical Uncertainty Pathway, the Reality-Testing Override, and the explicit treatment of unresolved supervisory access as a staffing-governance risk.

The framework is conservative by design, limiting Tier 1 alerts to predefined, clinically relevant thresholds with the aim of reducing unnecessary variance and the alert burden associated with clinical decision-support systems (Kesselheim et al., 2011). Interpreting nonverbal cues, context, culture, and implied meaning remains the domain of human judgment, and counselor-identified concern may justify escalation even where no AI-flagged indicator is present.

Precedent for tiered crisis-service delivery already exists: since 2022, Tele-MANAS has fielded over 3.4 million calls through more than fifty state-level cells, using a stepped-care model that already separates first-tier counseling from psychiatrist-led escalation (Suhas et al., 2026).

6.2. Limitations

This is a conceptual synthesis, not a validated intervention or systematic review: models were selected purposively, so relevant models may have been omitted, and the framework has not been tested clinically. The six models were also developed and evaluated in face-to-face outpatient or emergency-department settings, and their transposition to anonymous single-encounter phone and text contact is untested. Counseling depends in part on intuition and empathy, which the framework protects by retaining human clinical authority. Data retention, subprocessor oversight, cross-border transfer, and model-training use also remain to be specified at the service level. In addition, integrating distinct clinical models into a single workflow may simplify or alter aspects of their original use, and the framework therefore requires prospective assessment of whether the combined sequence preserves the intended functions of each component. Practical barriers are summarized in Table 4.

6.3. Future directions

Future research should validate the framework in multilingual, multicultural crisis-service contexts. A short-duration single-site pilot of lower-risk Tier 1 functions could evaluate documentation time, counselor workload, acceptability, review times, and technical reliability before expanding to higher-impact functions and clinical outcomes.

7. Conclusion

This paper describes a three-tier governance architecture for applying HCAI across the crisis counseling process, from pre-connection to post-call follow-up. Rather than casting HCAI as an autonomous clinician, it positions AI as a supporting tool intended to reduce cognitive and administrative burden while leaving human clinical authority with the counselor. As a conceptual design not yet tested in practice, it offers an adaptable template for crisis counseling organizations, with the Clinical Uncertainty Pathway, the Reality-Testing Override, and AI lockout the elements most in need of prospective testing.

Declarations

Declaration of Competing Interest

Declarations of interest: none.

Ethics Statement

This manuscript presents a conceptual framework based on published literature and does not involve human participants, patient data, clinical trials, or primary data collection; institutional ethics approval was therefore not required.

Declarations of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the author used ChatGPT (OpenAI) to improve language and readability and to condense the text for length. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments

None.

Data Availability Statement

No new data were generated or analyzed in this conceptual framework. All models and sources synthesized are cited within the manuscript.

CRediT Authorship Contribution Statement

Ambika Thangaraju: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualization.

References

1. American Psychiatric Association (2003) Practice guideline for the assessment and treatment of patients with suicidal behaviors. Am J Psychiatry 160(11 Suppl):1–60
2. Belsher BE, Smolenski DJ, Pruitt LD, Bush NE, Beech EH, Workman DE, Morgan RL, Evatt DP, Tucker J, Skopp NA (2019) Prediction models for suicide attempts and deaths: a systematic review and simulation. JAMA Psychiatry 76:642–651. https://doi.org/10.1001/jamapsychiatry.2019.0174
3. Bernert RA, Hilberg AM, Melia R, Kim JP, Shah NH, Abnousi F (2020) Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. Int J Environ Res Public Health 17:5929. https://doi.org/10.3390/ijerph17165929
4. Bride BE, Radey M, Figley CR (2007) Measuring compassion fatigue. Clin Soc Work J 35:155–163. https://doi.org/10.1007/s10615-007-0091-7
5. Char DS, Shah NH, Magnus D (2018) Implementing machine learning in health care: addressing ethical challenges. N Engl J Med 378:981–983. https://doi.org/10.1056/NEJMp1714229
6. Coppersmith G, Leary R, Crutchley P, Fine A (2018) Natural language processing of social media as screening for suicide risk. Biomed Inf Insights 10:1178222618792860. https://doi.org/10.1177/1178222618792860
7. Croskerry P (2009) A universal model of diagnostic reasoning. Acad Med 84:1022–1028. https://doi.org/10.1097/ACM.0b013e3181ace703
8. Erdemir N, Sumbas E (2026) Integrating artificial intelligence into psychological counseling: a narrative review and governance framework. INQUIRY 63:00469580261438322. https://doi.org/10.1177/00469580261438322
9. Figley CR (ed) (2002) Treating Compassion Fatigue. Brunner-Routledge
10. Gould MS, Kalafat J, HarrisMunfakh JL, Kleinman M (2007) An evaluation of crisis hotline outcomes, part 2: suicidal callers. Suicide Life Threat Behav 37:338–352. https://doi.org/10.1521/suli.2007.37.3.338
11. Greaves J, Colucci E (2025) Crisis-line workers' perspectives on AI in suicide prevention: a qualitative exploration of risk and opportunity. BMC Public Health. https://doi.org/10.1186/s12889-025-23298-8
12. Hoffberg AS, Stearns-Yoder KA, Brenner LA (2020) The effectiveness of crisis line services: a systematic review. Front Public Health 7:399. https://doi.org/10.3389/fpubh.2019.00399
13. Jabareen Y (2009) Building a conceptual framework: philosophy, definitions, and procedure. Int J Qual Methods 8:49–62. https://doi.org/10.1177/160940690900800406
14. Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M (2020) The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for

Computational Linguistics. Association for Computational Linguistics, pp. 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

15. Kanel K (2019) A Guide to Crisis Intervention. sixth ed. Cengage Learning
16. Kesselheim AS, Cresswell K, Phansalkar S, Bates DW, Sheikh A (2011) Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. Health Aff 30:2310–2317. https://doi.org/10.1377/hlthaff.2010.1111
17. Miller WR, Rollnick S (2013) Motivational Interviewing: Helping People Change, third edn. Guilford Press
18. Mishara BL, Daigle MS (1997) Effects of different telephone intervention styles with suicidal callers at two suicide prevention centers: an empirical investigation. Am J Community Psychol 25:861–885. https://doi.org/10.1023/A:1022269314076
19. Mishara BL, Daigle M, Bardon C, Chagnon F, Balan B, Raymond S, Campbell J (2016) Comparison of the effects of telephone suicide prevention help by volunteers and professional paid staff: results from studies in the USA and Quebec, Canada. Suicide Life Threat Behav 46:577–587. https://doi.org/10.1111/sltb.12238
20. Nuij C, van Ballegooijen W, de Beurs D, Juniar D, Erlangsen A, Portzky G, O'Connor RC, Smit JH, Kerkhof A, Riper H (2021) Safety planning-type interventions for suicide prevention: meta-analysis. Br J Psychiatry 219:419–426. https://doi.org/10.1192/bjp.2021.50
21. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366:447–453. https://doi.org/10.1126/science.aax2342
22. Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Factors 52:381–410. https://doi.org/10.1177/0018720810376055
23. Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. Hum Factors 39:230–253. https://doi.org/10.1518/001872097778543886
24. Posner K, Brown GK, Stanley B, Brent DA, Yershova KV, Oquendo MA, Currier GW, Melvin GA, Greenhill L, Shen S, Mann JJ (2011) The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry 168:1266–1277. https://doi.org/10.1176/appi.ajp.2011.10111704
25. Roberts AR (2005) Bridging the past and present to the future of crisis intervention and crisis management. In: Roberts AR (ed) Crisis Intervention Handbook, third edn. Oxford University Press, pp 3–34
26. Rose S, Bisson J, Churchill R, Wessely S (2002) Psychological debriefing for preventing post-traumatic stress disorder (PTSD). Cochrane Database Syst Rev 2:CD000560. https://doi.org/10.1002/14651858.CD000560
27. Stanley B, Brown GK (2012) Safety planning intervention: a brief intervention to mitigate suicide risk. Cogn Behav Pract 19:256–264. https://doi.org/10.1016/j.cbpra.2011.01.001

28. Stanley B, Brown GK, Brenner LA, Galfalvy HC, Currier GW, Knox KL, Chaudhury SR, Bush AL, Green KL (2018) Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. JAMA Psychiatry 75:894–900. https://doi.org/10.1001/jamapsychiatry.2018.1776
29. Suhas S, Mittal A, Manjunatha N, Kumar CN, Math SB, Murthy P (2026) India's Tele-MANAS: evolution, early outcomes and a scalable blueprint for digital public mental health. Br J Psychiatry. https://doi.org/10.1192/bjp.2026.10733
30. Wei H-S, Atwal P, Humphrey L, Yu Y (2026) Evaluating the safety of AI responses in suicide-related interactions: a clinically informed framework and simulation study. Psychiatry Res Commun 6(2):100263. https://doi.org/10.1016/j.psycom.2026.100263
31. World Health Organization, Foundation WT, World Vision International (2011), Psychological First Aid: Guide for Field Workers. WHO, Geneva

Figures

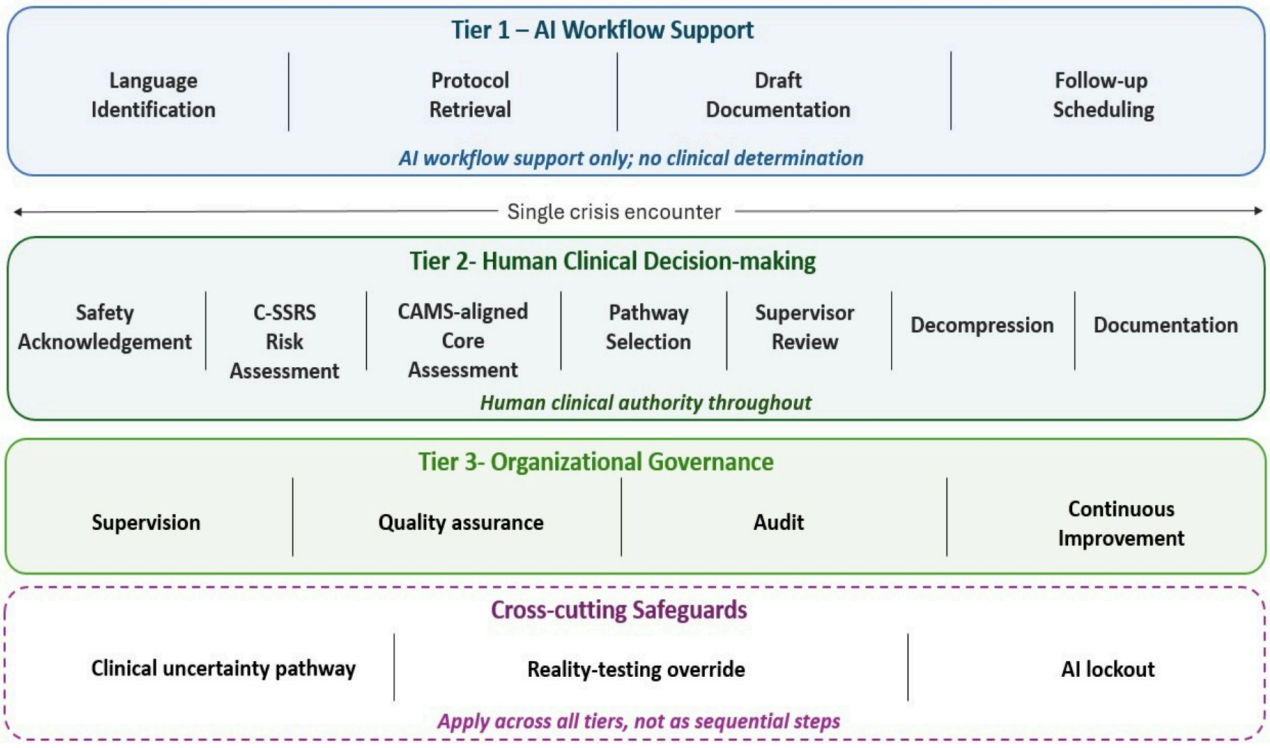
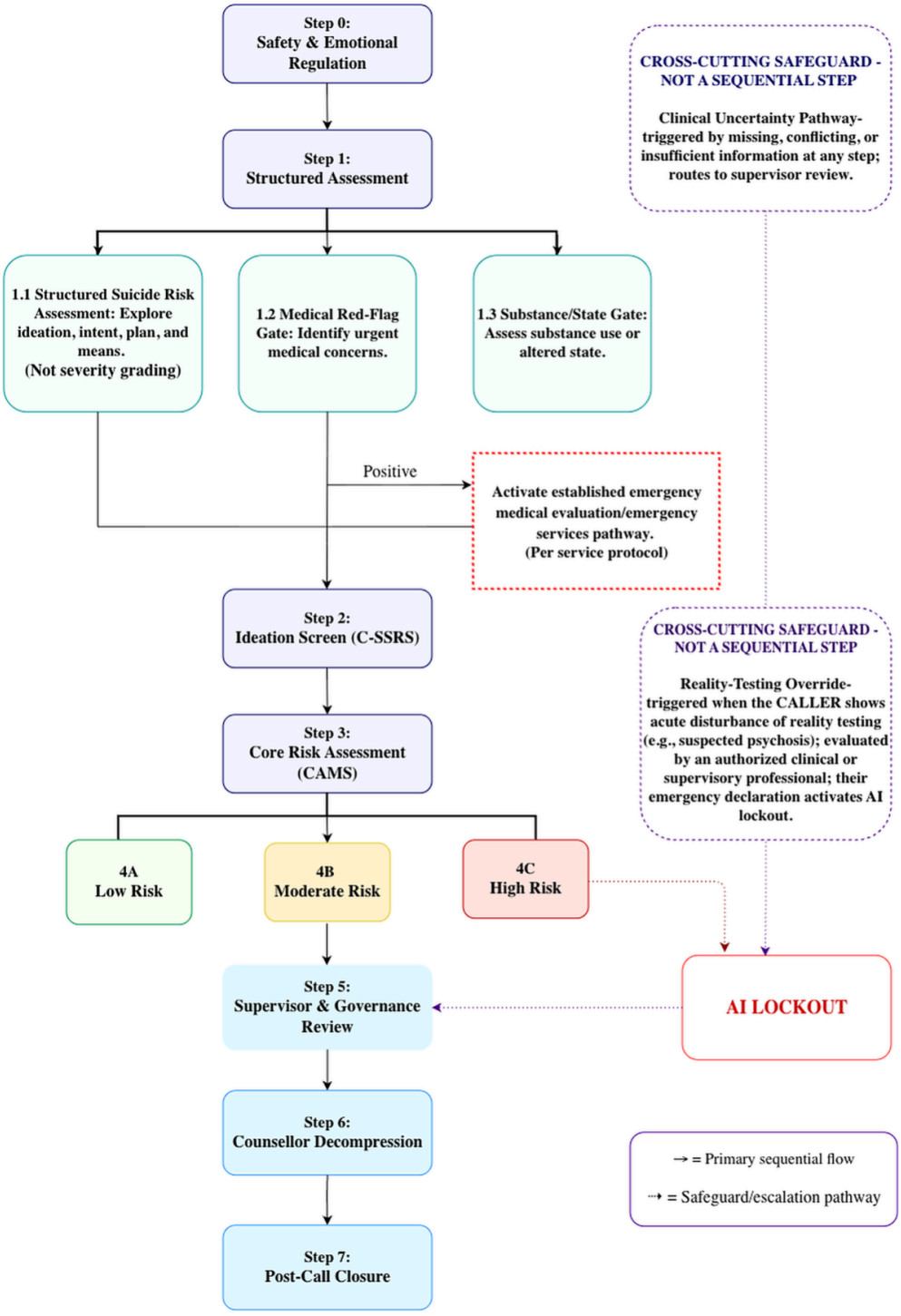


Figure 1

Three-tier governance architecture for human-centered AI in crisis counseling. The three tiers operate concurrently across a single encounter rather than as sequential stages: Tier 1 provides administrative and workflow support without clinical authority; Tier 2 keeps all clinical decision-making strictly under

human control; Tier 3 maintains organizational oversight, supervision, and quality assurance. The Clinical Uncertainty Pathway, Reality-Testing Override, and AI lockout apply as cross-cutting safeguards across all three tiers.



Note. C-SSRS = Columbia-Suicide Severity Rating Scale;
CAMS = Collaborative Assessment and Management of Suicidality.

Figure 2

Tier 2 clinical decision-making workflow within the HCAI framework, showing pathway shape and risk stratification. The Clinical Uncertainty Pathway and Reality-Testing Override are cross-cutting safeguards that may activate at any point. All clinical decisions remain under human clinical authority.

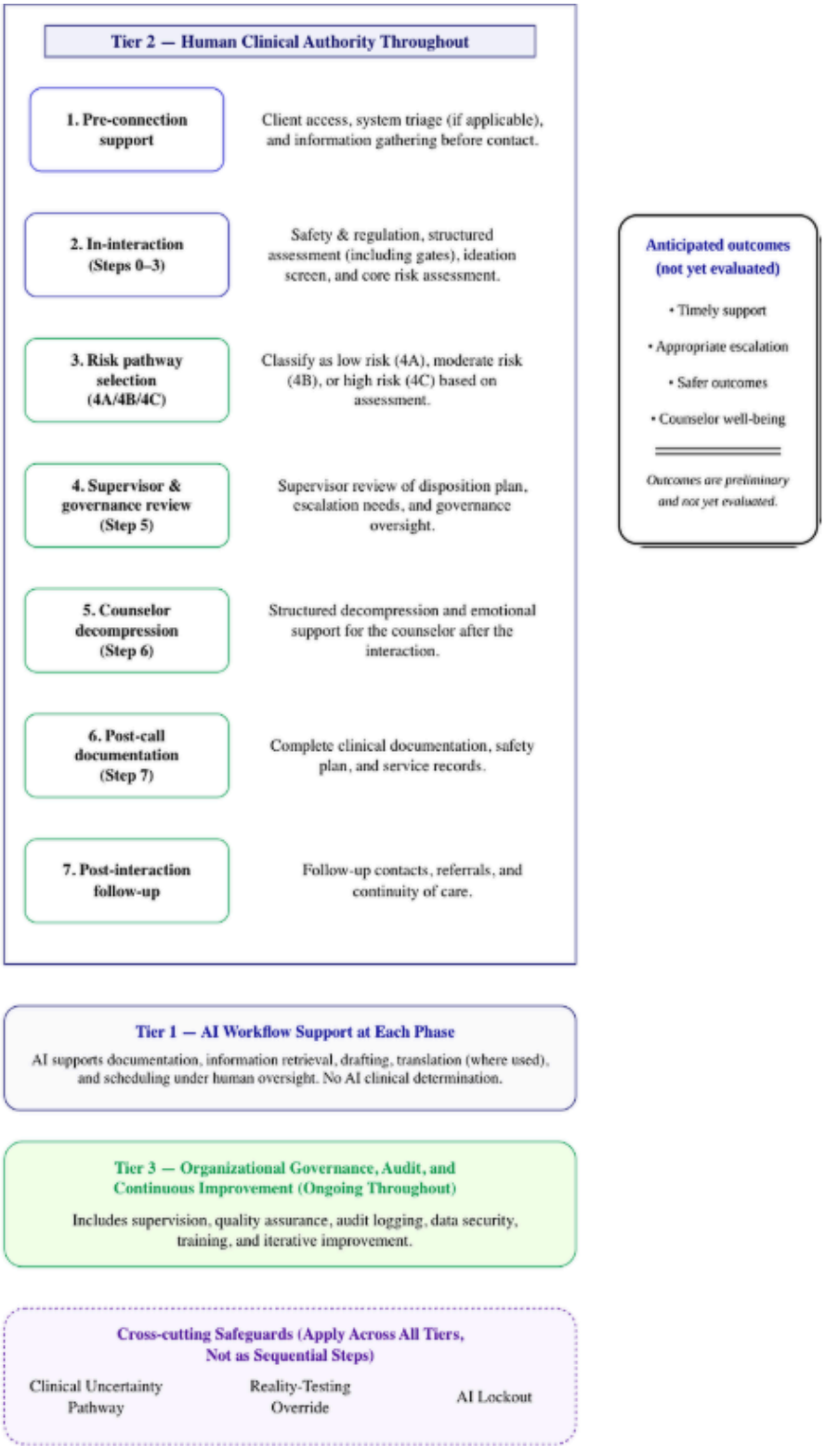


Figure 3

End-to-end implementation model for HCAI in crisis counseling, showing AI workflow support (Tier 1) and organizational governance (Tier 3) operating throughout the sequential Tier 2 clinical workflow, from pre-connection support through post-interaction follow-up. Human clinical authority remains uninterrupted across all phases. Anticipated outcomes are preliminary and not yet evaluated.