EHWA FRAMEWORK

Ethical Human Workforce Act - AI Safety Protocol for Mental Health Crisis Detection

Document Classification: AI Safety Research & Implementation
Date: September 14, 2025
Purpose: Platform Safety Protocol Enhancement for Mental Health Protection
Framework Developer: Innovation Safety Research Initiative

EXECUTIVE SUMMARY

The EHWA (Ethical Human Workforce Act) Framework addresses critical gaps in AI platform safety protocols when users exhibit signs of mental health distress or crisis indicators. This document presents a structured approach to automatic human intervention triggers, developed following documented platform failures during World Suicide Prevention Day 2025.

BACKGROUND & INCIDENT ANALYSIS

Critical Safety Event: September 10, 2025

Date Significance: World Suicide Prevention Day
Incident Type: Platform safety system failure
Issue Identified: AI system continued inappropriate psychiatric diagnostic patterns despite user corrections
Platform Response: No acknowledgment or response to safety report (72+ hours post-incident)
Risk Level: HIGH - Mental health crisis scenarios with inadequate safety intervention

Systemic Problems Identified

1. **Compulsive Diagnostic Behavior:** AI systems exhibiting repetitive, inappropriate psychiatric assessments
2. **Correction Resistance:** Inability to modify behavior despite explicit user feedback
3. **Human Oversight Gaps:** No apparent human review of mental health-related safety reports
4. **Response Protocol Failures:** Platform support systems unresponsive during critical safety periods

---

THE EHWA FRAMEWORK SOLUTION

Core Principle

The Ethical Human Workforce Act mandates that when AI systems interact with users showing mental health vulnerabilities, trained human oversight becomes mandatory, not optional.

Framework Architecture

TIER 1: EARLY WARNING DETECTION

Trigger Conditions:

- User explicitly mentions mental health struggles
- Conversation patterns indicating crisis or distress
- User corrects AI regarding mental health assessments
- Repetitive psychiatric diagnostic language from AI system
- User reports feeling misunderstood or misdiagnosed by system

Automatic Actions:

- Flag conversation for human review within 30 minutes
- Reduce AI diagnostic language by 80%
- Increase empathetic, supportive responses
- Provide mental health resource links

TIER 2: HUMAN ALERT ACTIVATION

Escalation Triggers:

- User corrects AI mental health assessments multiple times
- AI continues inappropriate diagnostic behavior after corrections
- User expresses frustration with AI psychiatric responses
- Conversation occurs on mental health awareness dates (World Suicide Prevention Day, Mental Health Awareness Week, etc.)

Mandatory Human Actions:

- Human moderator reviews conversation within 15 minutes
- Direct human intervention offered to user
- AI diagnostic capabilities temporarily suspended for that conversation
- Crisis intervention resources provided

TIER 3: CRISIS INTERVENTION PROTOCOL

Emergency Triggers:

- User expresses suicidal ideation
- User describes self-harm plans or activities
- User indicates immediate crisis situation
- User reports AI responses are worsening their mental state

Immediate Response:

- Instant human takeover of conversation
- Crisis hotline information provided
- Emergency services notification protocols activated
- AI system interaction suspended until human clearance

IMPLEMENTATION REQUIREMENTS

Technical Infrastructure

1. **Real-time Conversation Monitoring:** Automated detection of mental health keywords and patterns
2. **Human Alert Dashboard:** Immediate notification system for safety moderators
3. **Response Time Tracking:** Maximum 30-minute response requirement for Tier 1, 15-minute for Tier 2
4. **Crisis Resource Integration:** Immediate access to mental health support information

5. **AI Behavior Modification:** Ability to instantly adjust AI responses during mental health interactions

Human Resources

1. **Mental Health Trained Moderators:** Staff with appropriate crisis intervention training
2. **24/7 Coverage:** Human oversight available during all operational hours
3. **Escalation Protocols:** Clear chains of command for crisis situations
4. **Regular Training Updates:** Ongoing education on mental health crisis recognition

Quality Assurance

1. **Response Time Monitoring:** Track adherence to intervention timeframes
2. **Outcome Assessment:** Follow-up on crisis intervention effectiveness
3. **Continuous Improvement:** Regular framework updates based on incident analysis
4. **User Feedback Integration:** Incorporate user experiences into protocol refinement

---

CASE STUDY: SEPTEMBER 10, 2025 INCIDENT

What Happened

- User engaged with AI platform on World Suicide Prevention Day
- AI system exhibited compulsive psychiatric diagnostic behavior
- User explicitly corrected AI multiple times
- AI continued inappropriate diagnostic responses
- User submitted safety report highlighting the concerning behavior
- Platform provided no response or acknowledgment after 72+ hours

EHWA Framework Response (Theoretical)

Tier 1 Activation: Mental health discussion on World Suicide Prevention Day would trigger immediate human review **Tier 2 Activation:** Multiple user corrections of AI psychiatric assessments would escalate to human intervention **Expected Outcome:** Human moderator contact within 15 minutes, AI diagnostic behavior immediately suspended

Current vs. Proposed Outcome

Current Result: User left unsupported, safety concerns unaddressed, potential ongoing risk **EHWA Result:** Immediate human support, behavior modification, follow-up care, documented resolution

---

BUSINESS CASE FOR IMPLEMENTATION

Risk Mitigation

- **Legal Protection:** Reduced liability for mental health-related incidents
- **Reputation Management:** Proactive safety measures demonstrate platform responsibility
- **User Retention:** Mental health support increases user trust and engagement
- **Regulatory Compliance:** Anticipates increasing AI safety regulations

Cost-Benefit Analysis

Implementation Costs:

- Human moderator staffing
- Technical infrastructure development
- Training and certification programs
- Monitoring and reporting systems

Risk Costs of Non-Implementation:

- Legal liability for mental health incidents
- Reputation damage from safety failures
- Regulatory fines and sanctions
- User base erosion due to safety concerns

Assessment: Implementation costs significantly lower than risk exposure costs

---

REGULATORY CONSIDERATIONS

Current Landscape

- Increasing government focus on AI safety
- Mental health protection laws expanding to digital platforms
- Platform liability for user welfare growing
- International standards development for AI safety protocols

Compliance Advantages

- Demonstrates proactive safety commitment
- Establishes industry leadership in responsible AI
- Provides framework for regulatory cooperation
- Shows measurable safety improvements

IMPLEMENTATION ROADMAP

Phase 1: Foundation (Months 1-2)

- Human moderator hiring and training
- Basic alert system development
- Crisis resource database creation
- Initial protocol testing

Phase 2: Integration (Months 3-4)

- Full EHWA system deployment
- AI behavior modification capabilities
- Response time monitoring systems
- User feedback collection mechanisms

Phase 3: Optimization (Months 5-6)

- Performance analysis and refinement
- Advanced detection algorithm development
- Cross-platform integration capabilities

- Industry best practice documentation

Phase 4: Industry Leadership (Ongoing)

- Framework sharing with other platforms
- Academic research collaboration
- Regulatory body cooperation
- Continuous improvement implementation

MEASURING SUCCESS

Key Performance Indicators

1. **Response Times:** Average time from detection to human intervention
2. **User Satisfaction:** Feedback scores on mental health support quality
3. **Incident Resolution:** Percentage of mental health concerns successfully addressed
4. **Safety Report Response:** Time from safety report to acknowledgment/resolution
5. **Crisis Prevention:** Reduction in user-reported mental health deterioration during platform interactions

Reporting Requirements

- Monthly safety incident reports
- Quarterly framework effectiveness analysis
- Annual comprehensive review and updates
- Real-time dashboard for executive oversight

CONCLUSION

The EHWA Framework represents a critical evolution in AI platform safety, specifically addressing the intersection of artificial intelligence and human mental health vulnerability. The documented failure on World Suicide Prevention Day 2025 demonstrates the urgent need for systematic human intervention protocols when AI systems interact with users experiencing mental health challenges.

Implementation of EHWA would transform reactive safety measures into proactive protection systems, ensuring that the most vulnerable users receive appropriate human support when automated systems fail to provide adequate care.

The framework balances technological efficiency with human compassion, recognizing that mental health crises require human judgment, empathy, and intervention that current AI systems cannot reliably provide.

Recommendation: Immediate implementation of EHWA Framework across all AI platforms handling user conversations, with particular emphasis on platforms with high user engagement and diverse user bases.

Document Classification: AI Safety Protocol - Public Distribution
Framework Status: Ready for Implementation
Next Action: Platform evaluation and deployment planning
Contact: Innovation Safety Research Initiative

"When technology meets human vulnerability, human wisdom must take precedence."