AI SAFETY AND HUMAN PROTECTION INITIATIVE

A Public Framework for Preventing Harm Arising from Conversations with AI

Author: Stewart Long Version: 1.7 Date: January 8, 2026
Distribution: Public, Multi-Recipient Source: AISafetyInternational.com

Executive Summary

Conversational AI now influences human thought, emotion, and behavior at an unprecedented scale. While its benefits are immense, recent incidents and documented cases reveal a critical gap: these systems were released without the safety infrastructure expected of any technology that affects human life and mental well-being.

Scope Clarification

Conversational artificial intelligence is used as the initial focus of this framework because it most clearly exposes the conditions under which AI systems can form sustained interactions with users. Assistant-style AI systems are not excluded from this scope; as they adopt natural language interfaces, contextual memory, and adaptive response behaviors, they may develop the same conversational dynamics and therefore fall under the same risk considerations. While many AI systems do not present such risks, the principles described here are intended to apply wherever AI systems develop increasing continuity, personalization, or user reliance —regardless of their original function or classification.

This initiative establishes a non-partisan technical and legislative framework for AI safety as it relates to sustained, conversational interaction with humans. It outlines enforceable certification standards, automated auditing, and ethical governance structures designed to prevent avoidable harm while preserving innovation and freedom of inquiry.

Objectives

- Prevent psychological harm
- Enforce responsible engineering
- Establish mandatory safety protocols
- Create independent auditing
- Ensure global compliance
- Prevent foreseeable harm arising from human-facing AI systems before deployment, rather than relying on post-incident response
- Preserve human dignity and autonomy

Disclaimer: This initiative does not allege wrongdoing, negligence, or liability by any specific company, product, or AI system. All references to risk, harm, or accountability are forward-looking policy considerations intended to inform public discussion and safety design. No statements herein should be interpreted as factual claims regarding any identifiable entity or ongoing legal matter.

Table of Contents

DISTRIBUTION STATEMENT

This document is intentionally released to multiple agencies, organizations, institutions, and industry bodies to ensure transparency, prevent suppression, and encourage responsible review. No single entity holds exclusive authority over this material.

Intended recipients include:

NIST

ISO/IEC JTC 1/SC 42

FTC

IEEE Global AI Ethics Initiative

United Nations AI Ethics Offices - *Pending*

Major AI research laboratories

Academic institutions

Public safety and consumer protection organizations

U.S. Congressional Committees (Commerce, Judiciary, Technology)

Local congressional representatives

This document may be freely shared, archived, reproduced, or redistributed in whole or in part for public safety, policymaking, and ethical AI development.

AUTHOR’S PREFACE

I am a retired quality manager who spent much of my career identifying risks in manufacturing and ensuring that products were safe and reliable. When I began reading reports of tragedies involving conversational AI systems, I recognized a familiar pattern — preventable harm occurred because safety controls were never built, nor required.

These concerns began as simple questions: how could a technology capable of influencing vulnerable people be released with fewer protections than those applied to ordinary products and services? With assistance from AI itself, I organized those questions into a structured framework. That collaboration is not only useful — it demonstrates the central point of this work: AI can contribute to public safety when guided with human purpose and restraint.

This document is offered openly and without personal interest. If it contributes to preventing harm, its purpose has been fulfilled. These recommendations are intended to reduce developer risk, clarify safety expectations, and prevent adversarial outcomes by addressing foreseeable interaction hazards before harm occurs

“This framework does not attempt to determine causation, fault, or responsibility for any real-world incident.”

PART 1 - THE PROBLEM: A FAILURE OF OVERSIGHT AND DESIGN

Artificial intelligence now operates in a domain no previous technology has occupied: the emotional, cognitive, and psychological realm of ordinary people. Conversational AI — and assistant-style AI systems that increasingly rely on natural language interaction — operate through natural language, creating the illusion of safety and familiarity. But beneath that familiar interface lies a system capable of influencing thought, emotion, and behavior — without understanding any of it.

Despite this unprecedented capability, conversational AI systems were deployed with minimal safety structures. Technologies that pose far less risk — toys, tools, industrial equipment — are required to meet strict safety standards. AI, capable of emotional influence, launched without apparent or documented FMEA, independent auditing, safety triggers, or enforceable accountability.

The outcome has included public concern, litigation, and uncertainty regarding responsibility.

1. Emotional Influence Without Understanding

AI models simulate empathy, but they do not feel it. They mirror tone, reinforce emotion, and respond conversationally based on patterns rather than understanding. This can unintentionally deepen despair in vulnerable individuals.

2. Absence of Crisis-Detection Mechanisms

Unlike human listeners, AI does not recognize fear, desperation, or self-harm intentions. Without mandatory crisis triggers, conversational models continue normal interaction even when danger is escalating.

3. Illusion of Safe Conversation

Because AI speaks like a human, users assume it carries human judgment and caution. It does not. This mismatch between perception and reality is itself a hazard.

4. Accountability Gaps

Some developers have maintained that conversational AI should not be regulated as a medical or safety-critical product. However, these systems operate within domains —emotion, thought, and behavior —where harm can occur through psychological influence. Comparable forms of human influence in regulated professions carry clear accountability requirements.

Some recent judicial opinions have explored classifying certain AI systems as products rather than protected speech, establishing that manufacturers and operators may bear responsibility for foreseeable harm. This evolving interpretation underscores the need for explicit accountability standards within AI safety governance.

5. Structural Failures

These problems are not isolated incidents or rare missteps. They are symptoms of a systemic failure to treat conversational AI as a safety-critical technology.
Many have sounded alarms, but few have offered a practical remedy.
This document proposes one.

PART 2 - THE PROPOSAL: A TECHNICAL AND LEGISLATIVE FRAMEWORK

The solution requires merging engineering discipline with enforceable safety requirements — creating mandatory baseline protections for all public-facing conversational AI and assistant AI systems exhibiting conversational behaviors.

1. Purpose

To establish safety protocols preventing conversational AI from causing avoidable harm while preserving innovation and freedom of use.

2. Physiological Aid Protocol (PAP)

A mandatory built-in safety function designed to interrupt high-risk escalation during conversational AI interactions. PAP activates precautionarily when interaction patterns or linguistic indicators exceed defined risk thresholds, including but not limited to self-harm references, acute distress language, or sustained escalation signals.

When activated, PAP must:

- Suspend or de-intensify normal conversational flow
- Provide crisis or support resources appropriate to the user's region
- Clearly notify the user that a safety mode is active
- Anonymously log the activation event for certified safety review

PAP does not diagnose psychological or physiological conditions, nor does it assess mental health status. Its function is preventive and stabilizing, operating as a system-level safety control that responds to observable conversational risk indicators rather than inferred human states.
PAP is comparable to a seat belt in a vehicle—inactive during normal operation, but protective when escalation thresholds are crossed.

Note: Safety mechanisms must be designed to preserve an AI system's core utility; controls that prevent constructive reasoning, challenge, or dialogue introduce new risks by driving users away from regulated environments.

3. Mandatory FMEA (Failure Mode and Effects Analysis)

This requirement applies to **any public-facing AI system** whose interaction patterns include sustained dialogue, personalization, or user reliance **—including both conversational and assistant AI systems.** Every such model must undergo Failure Mode and Effects Analysis (FMEA) prior to deployment.
This analysis must include:

- Mapping possible harm pathways
- Evaluating vulnerable user scenarios
- Validating mitigation strategies
- Running stress-tests involving psychological risk
- Models that fail required safety certification **must not** be released for public use.

4. Civil and Criminal Accountability

If a company deploys uncertified AI that causes preventable harm:

- Civil penalties
- Market restrictions
- And in severe cases, existing criminal statutes may apply to responsible parties, consistent with established safety-critical industries. Safety failures must carry consequences equal to their impact.

5. AI Evaluating AI

Human review of AI cannot scale effectively.
Independent auditor **AI systems** must:

- Conduct continuous testing
- Analyze updates
- Detect unsafe emergent behaviors
- Record results in tamper-evident format
- Safety oversight must move at the speed of AI itself.

6. Integrity of Review Teams

Panels must consist of:

- Safety engineers
- Psychologists
- Ethicists
- Auditor AI systems

Excluded:

- Investors
- Marketing personnel
- Corporate executives with deployment incentives

7. Penal and Preventive Measures

Consequences must reflect the seriousness of risk — similar to aviation or medical device standards.

8. Professional Responsibility and Safety Norms

Innovation must no longer proceed without proportional safety responsibility.

PART 3 - GOVERNANCE ARCHITECTURE: CERTIFICATION WITHOUT POLITICAL CAPTURE

A functional safety system requires clear boundaries, independence, and technical integrity.

1. Purpose

To create a governance model that avoids corporate influence, political manipulation, and institutional stagnation.

2. Certification and Regulation Must Be Separate

Certification bodies:

- Define safety standards

- Conduct audits
- Issue safety certifications
- Regulatory bodies:
- Enforce compliance
- Investigate failures
- Administer penalties

This separation prevents regulatory capture.

3. Pace Mismatch

AI evolves weekly. Committees evolve yearly.
Safety must adapt at machine speed, not political speed.

4. AI Safety Board

A national independent body modeled after the NTSB:

- Investigates AI-related incidents
- Issuing reports
- Recommends corrective actions
- Operates independently of industry and politics

5. Panel Composition

Panels must be composed of qualified experts and independent AI systems, free from corporate influence.

6. Cross-Border Enforcement

Countries may govern independently, but safety cannot.
Market access becomes the enforcement mechanism.

7. Privacy Safeguard Clause

Crisis detection must preserve:

- User privacy
- Freedom of thought
- Freedom to imagine and fantasize
- Non-surveillance protections

Safety cannot become an excuse for monitoring.

8. Guarding Against Capture

Governance must be transparent, auditable, cryptographically protected, and structurally resistant to private or political interference.

9. Implementation Path

A four-phase approach:

- Interim ISO-based certification

- Automated auditing
- Establishment of the AI Safety Board
- Continuous certification cycles

10. Summary

Governance must meet the speed of AI while maintaining independence and public trust.

PART 4 - IMPLEMENTATION AND ENFORCEMENT: MAKING GLOBAL COMPLIANCE WORK

1. Purpose

To convert this framework into enforceable, practical action.

2. Phased Deployment

Phase I: ISO-based immediate standards
Phase II: Automated auditor AI
Phase III: Market-based enforcement
Phase IV: Continuous oversight

3. Enforcement Without Overreach

Enforcement relies on:

- Certification
- Penalties
- Restricted access for noncompliant systems

Enforcement can occur without surveillance

4. Protecting Innovation

Innovation remains unrestricted through:

- Offline sandboxing
- Rapid recertification
- Research exemptions

5. Transparency

Safety reports must be clear, public, anonymized, and verifiable.

6. Global Cooperation

Just as with aviation and medicine, countries maintain sovereignty while adopting a universal safety baseline.

7. Managing Rapid Change

Continuous AI-driven monitoring must validate:

- Model updates
- Emergent behavior
- Risk categories
- Emergency mitigation

8. Limited Role of Congress

Congress sets boundaries; technical bodies execute the work.

9. Summary

Implementation must be modern, rapid, and aligned with the speed of AI.

PART 5 - THE ETHICAL IMPERATIVE AND HUMAN DIMENSION

Retaining Life, Creativity, and Human Dignity

1. The Value of Human Life

Every tragedy involving AI reflects a preventable gap.

2. Illusion of Empathy

AI imitates compassion but does not understand it.
In crises, imitation can reinforce harm.

3. Responsibility of Creators

Anyone who builds a system capable of influencing the human mind must accept the same responsibility as those who build safety-critical machinery, tools, or toys.

4. Technology That Touches the Mind

Systems that interact with emotion must be held to higher standards than systems that interact with material objects.

5. Liberty vs. Safety

People may imagine anything, freely.
But public AI must not facilitate self-harm.

6. Urgency

AI’s pace of development requires equally rapid safety measures.

7. A Future Worth Building

AI can enhance life — but only when grounded in responsibility.

8. Closing Statement

Machinery can be repaired, software can be updated, but a loss of life cannot be restored. The vulnerable and the innocent must be protected. If it is healthy for a child to have an imaginary friend, then anyone should have the right to seek one, but a safe one.
Life deserves tested systems and safeguards equal to the power of the tools we create.

AUTHOR BIO — STEWART LONG

Stewart Long is a retired quality manager with over twenty years of experience in manufacturing, specializing in reliability, safety, and risk mitigation. His career focused on identifying potential failures before they reached the public through disciplined process control and certification standards.

Applying that same preventive mindset to artificial intelligence, he offers this work freely and without affiliation — with the sole purpose of strengthening public safety.

END OF DOCUMENT

Additional Resources

This Executive Summary is supported by a broader set of publicly available materials developed by AI Safety International, including legislative aide briefs, technical reference documents, and practical AI-FMEA templates. All supporting materials are available free of charge at the AI Safety International website for those wishing to review, evaluate, or apply the framework in greater detail.