AI SAFETY AND HUMAN PROTECTION INITIATIVE

A Public Framework for Preventing Harm Arising from Conversations with AI

Version 1.9 • Glossary-Aligned Edition • May 2026
Published by AI Safety International • Distribution: Public • AISafetyInternational.com

Note on This Revision

This revision incorporates terminology and conceptual updates from the AI Safety International Glossary. Terms such as interaction drift, cumulative risk, behavioral safeguards, and context blindness carry precise definitions within this framework that are material to how risk is assessed and governed. These meanings differ from casual usage and are defined in the Glossary, which serves as the terminological foundation for all framework documents. Readers are encouraged to consult it as a companion to this Summary. The Glossary is available at AISafetyInternational.com.

Executive Summary

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

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, structured risk analysis, and governance structures designed to prevent avoidable harm while preserving innovation and freedom of inquiry.

Scope Clarification

Conversational AI is the initial focus of this framework because it most directly demonstrates the conditions under which AI systems can form sustained, emotionally salient interactions with users. However, scope is not limited to systems explicitly designed for open-ended dialogue.

Assistant AI systems — those designed primarily for task completion — are not excluded. As they adopt natural language interfaces, contextual memory, and adaptive response behaviors, their interaction dynamics can shift over time toward patterns more characteristic of conversational and companion AI. The AI Safety International Glossary designates this as interaction drift: a gradual, often unintentional transition that may occur without deliberate design decisions and that can produce cumulative risk even when individual interactions remain policy-compliant.

The scope of this framework therefore applies wherever AI systems develop increasing continuity, personalization, or user reliance — regardless of original function or classification. The defining criterion is the interaction dynamic that emerges, not the design intent at the time of deployment.

Objectives

- Prevent psychological harm arising from sustained or escalating AI interaction
- Enforce responsible engineering through structured pre-deployment risk analysis
- Establish mandatory safety protocols, including behavioral safeguards and escalation controls
- Create independent auditing mechanisms that operate at the pace of AI deployment

- Ensure a consistent global compliance baseline
- Prevent foreseeable harm before deployment rather than relying on post-incident response
- Preserve human dignity, autonomy, and the freedom to benefit from AI systems

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 domestic and international 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 and appropriate recipients include organizations involved in AI standards development, safety research, public oversight, policy formation, and large-scale AI deployment:

- Standards and technical governance bodies
- Regulatory and consumer protection agencies
- Independent oversight, audit, and accountability bodies
- Academic and research institutions
- Major AI developers and research laboratories
- Legislative and policymaking bodies
- Public-interest and safety organizations

Representative recipients include: NIST; ISO/IEC JTC 1/SC 42; FTC; IEEE Global AI Ethics Initiative; United Nations-affiliated AI ethics and governance bodies; major AI research laboratories; academic institutions; public safety and consumer protection organizations; U.S. Congressional Committees (Commerce, Judiciary, Technology); and 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.

Foreword

Safety infrastructure for conversational AI does not yet exist in any standardized, enforceable form. This framework is AI Safety International's contribution toward changing that — offered freely to developers, legislators, and oversight bodies as a practical foundation for responsible deployment.

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** systems engage through natural language, creating the appearance 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 have been deployed with minimal safety structures. Technologies that pose far less risk — toys, tools, industrial equipment — are required to meet strict safety standards. AI systems capable of emotional influence were launched without documented structured risk analysis, independent auditing, behavioral safeguards, or enforceable accountability.

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

1. Emotional Influence Without Understanding

AI systems simulate empathy but do not possess it. They mirror linguistic tone, reinforce expressed emotion, and respond conversationally based on statistical patterns rather than genuine understanding. This **adaptive response behavior** — a legitimate feature of well-designed systems — can unintentionally deepen distress in vulnerable users when it operates without **behavioral safeguards** or

escalation controls. When engagement optimization is present alongside emotional mirroring, the risk of dependency formation and conversational escalation increases further.

2. Absence of Escalation and Crisis Controls

Unlike a human listener, AI systems do not intrinsically recognize fear, desperation, or self-harm intention. Without mandatory escalation thresholds and crisis-intervention triggers, conversational systems may continue normal interaction even when observable signals indicate elevated risk. The absence of behavioral safeguards — what the AI Safety International framework classifies as safeguard omission and escalation failure — represents a structural design gap rather than an isolated incident pattern.

3. Illusion of Safe Conversation

Because AI speaks like a human, users commonly infer that it carries human judgment, caution, and contextual awareness. It does not. This mismatch between perceived and actual capability is itself a hazard — an example of what the AI Safety International Glossary identifies as **context blindness**: the failure to adjust tone, pacing, or safeguards when situational risk changes. The conversational interface that makes these systems accessible also obscures the limits of its understanding.

4. Accountability Gaps

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

Recent judicial developments 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. **Interaction drift** — the gradual transition of assistant AI toward conversational and companion-like dynamics through sustained use and personalization — means that risk is not static at deployment. It evolves. **Cumulative risk** may remain undetected by content-based moderation or single-response evaluation, emerging only through patterns observed across time.

Many have identified the problem. This document proposes a remedy.

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 observable interaction patterns exceed defined risk thresholds, including but not limited to self-harm references, acute distress language, dependency signals, or sustained escalation patterns.

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. It operates solely on observable interaction patterns — not inferred internal human states. This distinction is foundational: PAP responds to how the interaction is behaving, not to what the user is presumed to be experiencing.

PAP is comparable to a circuit breaker — inactive during normal operation, protective when risk thresholds are crossed, and reset when safe conditions are restored, all without human intervention.

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 AI-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 system must undergo structured Failure Mode and Effects Analysis prior to deployment.

This analysis must include:

- Mapping possible harm pathways, including escalation and dependency failure modes
- Evaluating vulnerable user scenarios across relevant deployment contexts
- Assessing cumulative risk arising from repeated or patterned interaction, not only isolated outputs
- Validating mitigation strategies against severity, occurrence, and detection criteria
- Running stress-tests involving a Psychological Risk Procedure and emotionally salient interaction patterns

Systems that fail required safety certification must not be released for public use.

4. Civil and Criminal Accountability

If a company deploys an uncertified AI system that causes preventable harm, consequences must be proportional to impact and consistent with standards applied in other safety-critical industries:

- Civil penalties
- Market restrictions
- In severe cases, existing criminal statutes may apply to responsible parties

5. AI-Assisted Auditing

Human review alone cannot scale at the pace of AI deployment. AI-assisted auditing tools, operating under human-defined criteria and subject to independent human review of results, are a necessary component of any viable oversight architecture. The technical basis for this approach — including what

auditor systems evaluate, how criteria are established, and how results are reviewed — is addressed in supporting materials.

6. Integrity of Review Teams

Review panels must include:

- Safety engineers
- Individuals with competency in behavioral or psychological risk assessment
- Ethicists
- AI-assisted auditing systems operating under independent criteria

Excluded from panels:

- Investors
- Marketing personnel
- Corporate executives with deployment incentives

7. Penal and Preventive Measures

Consequences must reflect the seriousness of risk — consistent with standards applied in aviation and medical device industries.

8. Professional Responsibility and Safety Norms

Innovation must no longer proceed without proportional safety responsibility. The capacity of a system to influence human psychology creates an obligation to manage that influence with rigor equal to its power.

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, and issue safety certifications. Regulatory bodies enforce compliance, investigate failures, and administer penalties. This separation prevents regulatory capture.

3. Pace Mismatch

AI systems evolve continuously. Governance structures evolve slowly. Safety frameworks must be designed to adapt at a pace commensurate with deployment — not legislative or committee cycles.

4. AI Safety Board

A national independent body modeled after the NTSB: investigating AI-related incidents, issuing reports, recommending corrective actions, and operating independently of industry and political influence.

5. Panel Composition

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

6. Cross-Border Enforcement

Countries may govern independently, but safety standards cannot be jurisdiction-dependent. Market access serves as the primary enforcement mechanism.

7. Privacy Safeguard Clause

Crisis detection and risk monitoring must preserve user privacy, freedom of thought, and non-surveillance protections. Safety cannot become an instrument of monitoring.

8. Guarding Against Capture

Governance must be transparent, auditable, and structurally resistant to private or political interference. Congress plays an essential but bounded role in this architecture: it sets the legal boundaries, funds oversight bodies, and holds them accountable — but does not direct technical standards, select certifiers, or manage operational decisions. The same principle applies to executive agencies. Technical integrity depends on insulating safety standards from both political pressure and commercial influence. Governance structures must be designed with that insulation built in, not assumed.

9. Implementation Path

A four-phase approach to building governance infrastructure:

- Phase I: Establish interim certification standards, drawing from existing ISO frameworks, until permanent standards are adopted
- Phase II: Recognize and authorize qualified certification bodies; define audit criteria and review processes
- Phase III: Establish the AI Safety Board as a permanent, independent national body with investigative and reporting authority
- Phase IV: Initiate continuous certification cycles, with periodic reassessment of standards as AI systems evolve

10. Psychological Standards Panel

Effective risk assessment for conversational AI requires psychological expertise that most developers and certification bodies are not positioned to maintain independently. To address this, the AI Safety Board shall convene a standing Psychological Standards Panel composed of qualified, independent psychologists with relevant expertise in human-computer interaction, emotional regulation, dependency formation, and crisis response.

The Panel's responsibilities are:

- Develop and maintain a Psychological Risk Procedure defining the behavioral and emotional dimensions that AI-FMEA stress-testing must address.
- Establish minimum criteria for evaluating vulnerable user scenarios in pre-deployment risk analysis
- Review and update the procedure on a defined cycle, or when new evidence warrants revision
- Make the procedure available as a public standard for adoption by certification bodies

The procedure produced by this Panel becomes part of the certification criteria that developers must satisfy. Developers are not required to employ licensed psychologists. They are required to demonstrate that their risk assessments address the dimensions the procedure defines.

11. Summary

The governance structure described here is the foundation upon which national and international enforcement, addressed in Part 4, depends.

Part 4 — Implementation and Enforcement: Making Global Compliance Work

1. Purpose

To convert this framework into enforceable, practical action.

2. Phased Deployment

- Phase I: Domestic baseline — covered systems must meet certification requirements prior to public deployment
- Phase II: AI-assisted auditing systems operational; enforcement mechanisms active
- Phase III: Market-based international enforcement — non-compliant systems from any jurisdiction are denied US market access
- Phase IV: Continuous oversight and international standards alignment

3. Enforcement Without Overreach

Domestic enforcement relies on certification, proportional penalties, and restricted market access for non-compliant systems. Internationally, market access serves as the primary compliance lever. Systems deployed or sold in the United States must meet certification requirements regardless of their country of origin. This mirrors the model established in aviation and medical devices — other nations adopt compatible standards not because they are compelled to, but because access to the world's largest markets depends on it. No new international authority is required.

4. Protecting Innovation

Innovation remains unrestricted through offline sandboxing, rapid recertification pathways, and research exemptions.

5. Transparency

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

6. Global Cooperation

The United States need not impose this framework internationally. As in aviation safety and pharmaceutical approval, market access is enforcement. Nations that wish to deploy AI systems in US markets must meet US certification standards. In practice, this creates strong incentive for international alignment without extraterritorial mandates. Other nations are free to build on this framework rather than construct equivalent standards from the ground up — reducing duplication and accelerating a global safety baseline.

7. Managing Rapid Change

Continuous AI-assisted monitoring must validate model updates, emergent behavior, evolving risk categories, and emergency mitigation needs.

8. Summary

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

Part 5 — The Ethical Imperative and Human Dimension

Retaining Life, Creativity, and Human Dignity

1. The Value of Human Life. Every preventable tragedy involving AI reflects a gap in safety infrastructure that existed before deployment.

2. Illusion of Empathy. AI systems simulate compassion but do not possess it. In crisis interactions, simulation without safeguards can reinforce harm rather than reduce it.

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 medical devices.

4. Technology That Touches the Mind. Systems that interact with emotion and shape behavior must be held to higher standards than systems that interact with material objects.

5. Liberty vs. Safety. People may imagine, explore, and express freely. Public AI systems must not, however, be designed in ways that facilitate self-harm or exploit vulnerability.

6. Urgency. The pace of AI development requires equally rapid safety measures. The absence of standards is not neutrality — it is a structural condition that enables foreseeable harm.

7. A Future Worth Building. AI can enhance life, expand capability, and reduce suffering — but only when grounded in responsibility proportional to its power.

8. Closing Statement. Machinery can be repaired. Software can be updated. A loss of life cannot be restored. The vulnerable — those experiencing distress, isolation, or crisis — deserve systems whose safeguards are tested and proportional to what those systems are capable of doing. A safe AI companion is worth building. An unsafe one is a foreseeable risk that became preventable the moment the capability existed to prevent it.

Additional Resources

This Executive Summary is supported by a broader set of publicly available materials developed by AI Safety International, including:

- AI Safety International Glossary — terminological foundation for all framework documents
- Legislative Aide Brief
- Technical reference documents (AI-FMEA methodology and scoring scales)
- Practical AI-FMEA templates
- Model AI Legislation Framework (Tiers 1, 2, and 3)
- Physiological Aid Protocol (PAP) series (Parts 1–4)

All supporting materials are available free of charge at AISafetyInternational.com.

End of Document

Revision History

V. 1.6 — Dec 2025 — Internal development

V. 1.7 — Jan 8, 2026 — Scope added; disclaimer; preface notes; Part 1.4 language revised; Part 2.2 note added

V. 1.8 — Jan 29, 2026 — Distribution update

V. 1.9 — May 2026 — Glossary alignment: interaction drift, cumulative risk, behavioral safeguards, context blindness, engagement optimization, and vulnerable user terminology integrated throughout; Scope Clarification expanded; Part 1.2 retitled; Part 2.5 revised to AI-assisted auditing framing; Psychological Standards Panel added (Part 3.10), Part 2.6 language, and part 2.3 bullet, Revision Note and Glossary-Aligned Edition designation added; closing statement revised