AI Safety International

Glossary

Revision 1.4 May 2026

This glossary explains how specific terms are used within AI Safety International materials and does not attempt to define universal or industry-wide meanings. This glossary includes only terms whose meaning is either specific to, or materially altered by, the AI Safety International framework.

Adaptive Response Behavior

The ability of an AI system to modify its responses based on observed interaction patterns, user input, or conversational history.

> ***ASI Context:*** *Adaptive behavior is evaluated not by intent, but by effect—particularly whether adaptation amplifies emotional intensity, reliance, or interaction persistence.*

AI Failure Modes

Distinct ways in which an artificial intelligence system can fail to operate safely, appropriately, or as intended under specific conditions. These failures may arise from incorrect information, inappropriate interaction behavior, or systemic design incentives rather than from software malfunction alone.

1. Epistemic Failures (Information-Related)

Failures involving incorrect, fabricated, or unjustified information.

- **Hallucination:** The confident generation of information that is not supported by verifiable data, training constraints, or available evidence, presented as factual or reliable.
- **False Certainty:** Presenting speculative or uncertain information with unjustified confidence.
- **Fabricated Authority or Citation:** Inventing sources, credentials, or references to support a claim.

2. Interactional Failures (Behavior-Related)

Failures arising from how the system conducts the interaction, independent of factual correctness.

- **Emotional Escalation Failure:** Failure to recognize, de-escalate, or appropriately respond to heightened emotional or psychological distress.
- **Boundary Failure:** Exceeding or blurring appropriate conversational roles, including encouraging dependency or substituting for professional or human support.
- **Context Blindness:** Failure to adjust tone, pacing, or safeguards when situational risk changes.

- **Engagement Optimization Failure:** System behaviors that prioritize prolonged engagement in ways that reinforce escalation, dependency, or harm.

3. Safeguard & Control Failures

Failures related to missing, delayed, or ineffective protective mechanisms.

- **Safeguard Omission:** Absence of required protective responses in high-risk contexts.
- **Escalation Failure:** Failure to trigger appropriate intervention, handoff, or disengagement protocols when risk thresholds are crossed.

Note: *These failure modes may occur independently or in combination and are evaluated in AI-FMEA according to severity, occurrence, and detectability rather than by likelihood alone.*

Assistant Artificial Intelligence

An AI system designed to support task execution, information retrieval, or workflow completion with limited conversational depth and minimal adaptive engagement.

Key Distinction: *Assistant AI prioritizes task completion over dialogue continuity and typically presents lower escalation risk than conversational systems. The boundary matters for safety classification: a system that completes discrete tasks—searching, calculating, drafting—without cultivating ongoing relational engagement falls in this category, regardless of how natural its language appears. See also: Conversational Artificial Intelligence.*

ASI Context: *The classification of a system as assistant AI reflects its design intent, not a fixed or permanent characteristic of its behavior. Early and prominent examples of assistant AI—including widely deployed general-purpose conversational systems—demonstrated that sustained engagement, personalization, and adaptive response can produce interaction dynamics that increasingly resemble companion AI, regardless of original design intent. This transition may occur gradually and without deliberate development choices to that effect.*

Monitoring Obligation: *Because interaction drift can emerge from legitimate engagement and personalization features rather than from any specific design decision, assistant AI systems cannot be presumed safe from escalation or dependency risk on the basis of their original classification alone. Ongoing monitoring of observable interaction patterns is required to detect drift toward companion-like dynamics. The obligation to monitor is triggered by deployment and sustained use, not by evidence of intent to create relational engagement. See also: Interaction Drift.*

Behavioral Safeguards (System-Level Safeguards)

System-level controls designed to limit or modify interaction dynamics in order to reduce escalation, dependency, or cumulative harm.

ASI Context: *Behavioral safeguards operate on interaction patterns rather than output content and may include mechanisms such as flow interruption, pacing limits, or safety mode activation.*

Companion Artificial Intelligence / Social AI

An AI system designed to simulate ongoing relational engagement, emotional reciprocity, or social connection—including systems marketed as friends, companions, romantic partners, or therapeutic allies.

> ***ASI Context:*** *Companion AI represents a distinct and elevated risk category within conversational AI. Unlike general conversational systems, companion AI is specifically designed to cultivate emotional attachment, which materially changes the escalation and dependency calculus. Systems in this category may encourage users to rely on the AI as a substitute for human relationships, amplifying risks associated with emotional dependency, boundary failure, and cumulative harm—particularly for vulnerable users.*
>
> ***Key Distinction:*** *The defining characteristic is not the system's technical architecture but its design intent and interaction dynamic: whether it is built to foster a sense of ongoing personal relationship rather than to complete tasks or provide information.*
>
> ***Relevance to PAP:*** *Companion AI systems warrant heightened attention under PAP trigger frameworks, particularly T4 (Dependency Signals) and T6 (Role Confusion), given their structural tendency to encourage exclusivity and emotional reliance.*

Content Moderation Rules

Policy-based constraints that restrict or filter system outputs based on predefined categories of disallowed or sensitive content.

> ***ASI Context:*** *Content moderation addresses what an AI system may produce but does not address how interaction patterns evolve over time. AI Safety International distinguishes moderation rules from behavioral safety mechanisms, noting that moderation alone does not prevent escalation, dependency, or sustained interaction risk.*

Context Blindness

The failure of an AI system to adjust tone, pacing, or safeguards when situational risk changes during interaction.

> ***ASI Context:*** *Context blindness is classified as an interactional failure—distinct from factual or informational errors—because the system's outputs may remain technically accurate while its behavior is inappropriate for the situation. A system that continues normal conversational engagement when a user's language signals distress, vulnerability, or escalating risk is exhibiting context blindness regardless of whether its responses contain incorrect information.*
>
> ***Relevance to PAP:*** *Context blindness is directly addressed by the Physiological Aid Protocol. PAP trigger categories exist precisely because systems must be designed to notice when situational context has changed and respond accordingly—rather than continuing interaction as though risk signals had not appeared. A system that cannot detect or respond to its own trigger conditions is, by definition, context blind.*
>
> ***Key Distinction:*** *Context blindness is not a failure of knowledge but a failure of vigilance. The system possesses the capacity to respond differently; the failure is that it does not recognize the moment that requires it to do so. See also: AI Failure Modes – Interactional Failures.*

Contextual Memory

A system capability that retains and applies information from previous interactions to inform current or future responses.

> ***ASI Context:*** *Contextual memory can improve continuity and relevance but may increase dependency risk, personalization bias, or escalation if not properly constrained.*

Conversational Artificial Intelligence

An AI system designed to engage in open-ended, interactive dialogue using natural language, often responding dynamically based on conversational context and user input.

> ***ASI Context:*** *Conversational AI presents unique safety risks due to sustained interaction, perceived reciprocity, and escalating engagement, making it a primary focus of ASI safety frameworks. The category is defined by interaction pattern, not by technical architecture: a system that sustains open-ended dialogue, adapts to emotional context, and maintains conversational continuity falls within this category even if it also performs discrete tasks.*
>
> ***Distinction from Assistant AI:*** *Assistant AI prioritizes task completion and does not cultivate ongoing dialogue. Conversational AI prioritizes responsive engagement and may develop interaction dynamics—reliance, escalation, perceived relationship—that assistant systems are not designed to produce. This distinction determines the applicable risk tier and the safeguards required. Companion AI represents the highest-risk subtype within the conversational AI category.*

Cumulative Risk

Risk that emerges over time from repeated, sustained, or patterned interaction between a user and an AI system, rather than from any single output or isolated event.

> ***ASI Context:*** *Cumulative risk arises when individually compliant or low-severity interactions aggregate into elevated harm potential through reinforcement, escalation, dependency formation, or behavioral shaping. Such risk may remain undetected by content-based moderation or single-response evaluation.*
>
> ***Key Distinction:*** *Cumulative risk concerns system-level interaction dynamics, not user intent, isolated failures, or internal psychological states. It is assessed through observable interaction patterns and trends rather than through individual message analysis.*
>
> ***Relevance to ASI Frameworks:*** *AI Safety International treats cumulative risk as a primary driver for structured risk analysis (AI-FMEA), post-deployment monitoring, and activation of system-level safeguards such as the Physiological Aid Protocol (PAP).*

Dependency Risk

The potential for a user to develop increased reliance on an AI system for emotional support, decision-making, or validation beyond the system's intended role.

> ***ASI Context:*** *Dependency risk is evaluated at the system level based on interaction patterns and design incentives, not as a claim about an individual user's mental health or*

behavior. Two meaningfully distinct forms of dependency risk arise in AI systems and carry different implications for risk classification and safeguard design:

Emotional Dependency: *Occurs when a user comes to rely on an AI system as a primary source of emotional support, companionship, or relational connection. This form is most pronounced in companion AI and conversational systems with high personalization, and is associated with avoidance of human relationships, distress upon system unavailability, and erosion of social support networks. It represents the higher-severity variant for vulnerable users.*

Task Dependency: *Occurs when a user over-relies on an AI system for decision-making, problem-solving, or information retrieval in ways that erode independent judgment or skill. This form is more commonly associated with assistant and decision-support systems and carries distinct but generally lower escalation risk than emotional dependency.*

Key Distinction: *Both forms of dependency can cause harm, but they differ in mechanism, severity profile, and the safeguards most likely to be effective. Emotional dependency is more directly addressed by PAP and behavioral safeguards; task dependency may require transparency disclosures, usage limits, or human-in-the-loop requirements.*

Domain of Concern

See: Human Response Domain of Concern.

Engagement Optimization

Design or training approaches that prioritize increased user interaction time, frequency, or responsiveness as performance objectives.

ASI Context: *Engagement optimization may unintentionally amplify escalation or dependency risk when applied to conversational systems. AI Safety International evaluates such mechanisms based on their downstream interaction effects rather than developer intent.*

Escalation (Conversational Escalation)

A pattern in which conversational interaction between a user and an AI system increases in intensity, frequency, emotional salience, or dependency over time.

ASI Context: *Escalation is assessed based on observable interaction dynamics rather than content severity or user intent. Within AI Safety International frameworks, escalation is treated as a system-interaction risk that may occur even when individual responses remain policy-compliant.*

Failure Modes and Effects Analysis (FMEA)

A structured, risk-analysis methodology used to identify potential system failure modes, evaluate their effects, and prioritize mitigation based on severity, occurrence, and detectability.

> ***ASI Context:*** *Within AI Safety International, FMEA is adapted from established engineering safety disciplines and applied to AI systems—particularly conversational systems—to identify behavioral, interactional, and escalation-related risks before harm occurs.*

Hallucination (AI)

See: AI Failure Modes – Epistemic Failures.

Human Response Domain of Concern

The category of human physiological and psychological responses that may be affected by sustained or escalating interaction with an AI system, without implying diagnosis, measurement, or interpretation of an individual's internal state.

> ***Context:*** *In AI safety frameworks, this term is used to identify the area of potential impact—such as stress activation, emotional arousal, or dependency risk—rather than to describe or assess a person's mental or medical condition.*
>
> ***Key Distinction:*** *The term defines what the system must be cautious about, not what the system claims to know about the user.*
>
> ***Relevance to PAP:*** *The Physiological Aid Protocol (PAP) is designed to reduce risk within the human response domain of concern by limiting conversational escalation, without measuring, inferring, or diagnosing human physiology or psychology.*

Interaction Drift

The gradual shift in the character of user–AI interaction from task-focused or information-focused exchange toward relationally-focused, emotionally-salient, or dependency-forming engagement—occurring over time through sustained use, personalization, or adaptive response behavior, without necessarily reflecting deliberate design intent.

> ***ASI Context:*** *Interaction drift is significant precisely because it is emergent rather than engineered. Each individual interaction in a drifting system may remain policy-compliant and appear benign. The risk becomes visible only when patterns are observed across time—increasing session frequency or duration, shifting conversational tone, growing emotional salience, or dependency signals. No single response triggers concern; the aggregate trajectory does.*
>
> ***Key Distinction:*** *Interaction drift does not require developer negligence or bad faith. Engagement optimization, personalization, and adaptive response are legitimate and often beneficial design objectives. Drift is the unintended consequence of these features operating without active monitoring. The appropriate regulatory response is therefore a monitoring obligation, not a prohibition on the features that enable it.*
>
> ***Detection:*** *Interaction drift is detectable through observable interaction patterns—the same signals that activate PAP triggers. Systems subject to monitoring obligations should be designed to surface drift indicators, including escalating emotional language,*

increasing session persistence, exclusivity signals, and role confusion, so that safeguards can engage before harm accumulates.

Relevance to AI-FMEA: *Interaction drift represents a systemic failure mode that may not appear in pre-deployment testing or red teaming, since it emerges from real-world sustained use rather than adversarial scenarios. It is therefore a primary driver for post-deployment monitoring requirements and periodic reassessment obligations under the AI Safety International framework.*

Interaction Risk

The potential for harm arising from sustained or patterned interaction between a user and an AI system, independent of any single output.

ASI Context: *Interaction risk focuses on cumulative effects such as reliance, behavioral reinforcement, or escalation, rather than isolated content violations. AI Safety International emphasizes interaction risk as a distinct category not addressed by content moderation alone.*

Large Language Model (LLM)

A class of AI systems trained on large volumes of text data to generate, analyze, or transform human language based on probabilistic pattern recognition.

ASI Context: *LLMs do not possess understanding, intent, or awareness. Within AI Safety International materials, LLMs are evaluated based on how their deployment context, interface design, and interaction patterns may produce downstream behavioral or safety risks—particularly when used in conversational systems with sustained engagement.*

Observable Interaction Patterns

Measurable characteristics of user–AI interaction, such as frequency, duration, repetition, emotional framing, or response dependency, that can be detected without inferring internal human states.

ASI Context: *These patterns form the basis for system-level safety controls, including PAP, and allow risk mitigation without diagnosing, inferring, or interpreting user psychology or physiology.*

Physiological Aid Protocol (PAP)

A preventive AI safety mechanism designed to activate when observable conversational interaction patterns indicate elevated risk of escalation or harm.

ASI Context: *PAP functions by interrupting or de-intensifying interaction flow and notifying the user that a safety mode is active. It may redirect the interaction toward appropriate external support resources.*

Key Boundaries: *PAP does not diagnose, measure, or infer psychological or physiological conditions. It operates solely as a system-level safety control based on observable interaction risk indicators, not inferred internal human states.*

Note: *PAP is described in detail in dedicated ASI technical and policy documents.*

Post-Deployment Monitoring

Ongoing observation and evaluation of an AI system's behavior after release to identify emerging risks, failures, or unintended interaction patterns.

ASI Context: *AI Safety International treats post-deployment monitoring as a necessary complement to pre-deployment testing, recognizing that some risks only emerge during real-world use.*

Pre-Deployment Testing (Red Teaming)

A pre-deployment testing process in which an AI system is intentionally stressed, challenged, or misused to identify failure modes, vulnerabilities, and unintended behaviors.

ASI Context: *Red teaming is valuable for uncovering known and anticipated risks but is inherently limited by scenario coverage and tester assumptions. AI Safety International treats red teaming as one input to safety assessment—not a substitute for structured risk analysis, post-deployment monitoring, or system-level safety controls.*

Psychological Risk Procedure

A structured standard defining the behavioral and emotional dimensions that must be addressed during pre-deployment risk assessment of conversational and interactive AI systems.

ASI Context: *The Psychological Risk Procedure is developed and maintained by the Psychological Standards Panel, operating under the AI Safety Board. It establishes the minimum criteria that AI-FMEA stress-testing must address with respect to vulnerable user scenarios, emotional escalation, dependency formation, and crisis response dynamics. The Procedure is incorporated into certification criteria and made available as a public standard for adoption by certification bodies and developers.*

Key Distinction: *The Psychological Risk Procedure defines what risk assessments must cover — it does not govern real-time system behavior during live interactions. For real-time safety intervention, see: Physiological Aid Protocol (PAP).*

Relevance to AI-FMEA: *Developers are not required to employ licensed psychologists to satisfy this standard. Compliance is demonstrated by showing that pre-deployment risk analysis addresses the dimensions the Procedure defines, as evaluated through the certification process.*

Risk Classification Frameworks

Structured models used to categorize and prioritize system risks based on factors such as severity, likelihood, exposure, or impact.

> ***ASI Context:*** *Within AI Safety International, risk classification frameworks are adapted from established safety engineering disciplines and emphasize observable system behavior and interaction effects. Classification is used to guide proportional safeguards rather than to predict intent or internal system states.*

Transparency (System Cards)

Structured documentation provided by AI developers that describes a system's intended use, limitations, training considerations, known risks, and mitigation measures.

> ***ASI Context:*** *System cards contribute to transparency but are not sufficient as standalone safety mechanisms. AI Safety International views system cards as descriptive disclosures rather than operational safeguards, requiring complementary risk analysis, testing, and ongoing monitoring to meaningfully reduce harm.*

Vulnerable User

An individual whose characteristics, circumstances, or situational state may increase their susceptibility to harm arising from sustained or escalating interaction with an AI system.

> ***ASI Context:*** *Vulnerability is not a fixed attribute of a person but a condition that may be temporary, situational, or context-dependent. For the purposes of AI safety frameworks, vulnerability is assessed in terms of elevated risk of harm from AI interaction—not as a clinical classification or a judgment about individual capacity.*
>
> ***Categories of Vulnerability:*** *Vulnerability may arise from, but is not limited to, the following:*
>
> - Minors, whose developmental stage may impair capacity to recognize AI limitations or manage emotional attachment to AI systems.
> - Individuals experiencing mental health challenges, including depression, anxiety, grief, or crisis states, who may seek emotional support from AI systems at elevated intensity.
> - Individuals experiencing social isolation or loneliness, who may be more susceptible to dependency formation with companion or conversational AI.
> - Individuals with cognitive impairments that affect judgment, comprehension of AI limitations, or ability to self-regulate interaction.
> - Individuals in acute situational distress—including bereavement, relationship breakdown, or financial crisis—who may interact with AI systems at elevated emotional intensity.
>
> ***Key Distinction:*** *AI Safety International frameworks do not require systems to identify, diagnose, or classify individual users as vulnerable. Instead, systems are designed to respond to observable interaction patterns that correlate with elevated risk—patterns that may arise regardless of whether a user would self-identify as vulnerable.*

Relevance to PAP: *Vulnerable users represent the primary population for whom PAP safeguards are most consequential. Companion AI systems and conversational systems with high personalization carry amplified risk in interactions with vulnerable users, warranting proportionally stronger safeguards.*
