AffixIO Technical Paper · WP-013
June 2026
affix-io.com

AFFIXIO WHITE PAPER · WP-013

AI Safety at the Infrastructure Layer: The Cryptographic Distress Guardrail

Respond to crisis signals without building a database of the worst things users say.

AffixIO | United Kingdom | affix-io.com | June 2026

ABSTRACT

Safety tooling usually means storing the conversations you most need to protect. AffixIO detects distress tiers at the policy enforcement layer, returns human-reviewed crisis resources, and logs only hashed safety events anchored in Merkle proofs.

CONTENTS

SECTION 1

Introduction

Conversational AI systems encounter users in distress. The statistical inevitability is clear: a large-scale AI service serving millions of interactions will encounter users expressing suicidal ideation, self-harm, domestic violence, child safeguarding concerns, and mental health crises. The question is not whether these situations will arise but what the AI system does when they do.

Most AI service providers address this through model-level training (the model is trained to recognise and respond appropriately to safety-critical content), content moderation (flagged conversations are reviewed by human moderators), and crisis resource surfacing (the AI includes crisis helpline information in responses to distress-indicating queries). All of these approaches have merit. They also share a common characteristic: they process the content of potentially sensitive conversations for safety purposes, creating retention obligations for the most sensitive material users share.

AffixIO's cryptographic distress guardrail implements safety monitoring at the infrastructure layer rather than the model layer. The policy enforcement layer, which already evaluates every AI response before delivery, includes a distress pattern detector that operates at the prompt and response level. When a distress pattern is detected, the safety response is triggered at the infrastructure layer before the AI response is delivered. The safety event is logged using a hashed session identifier; the content that triggered the

detection is not stored. The ZK-anchored safety event record proves that a safety response was triggered and delivered without revealing the triggering content.

SECTION 2

Regulatory Context

The Online Safety Act 2023 (UK) places safety duties on regulated user-to-user services and search services. Schedule 7 identifies categories of "primary priority harmful content" that services must take steps to prevent children from encountering. Schedule 11 identifies illegal content including suicide and self-harm communications. Section 12 requires that services with functionality allowing users to interact with others or receive AI-generated content implement systems to identify and address safety risks.

OFCOM's Online Safety Act guidance identifies "safety by design" as the preferred approach: safety mechanisms should be built into the service architecture, not applied as optional filters. The cryptographic distress guardrail is precisely a safety-by-design mechanism: it operates at the infrastructure layer, cannot be bypassed by users, and generates a verifiable record of safety responses.

The EU AI Act classifies AI systems intended to interact with natural persons as potentially high-risk where they may cause harm to vulnerable persons. Annex III (high-risk AI systems) includes AI systems used in education, employment, and essential services affecting vulnerable groups. Article 9 requires that high-risk AI system providers implement risk management systems that address foreseeable misuse. Distress detection and safety response is a direct implementation of Article 9's foreseeable misuse risk management requirement.

SECTION 3

The Privacy-Safety Tension

Safety monitoring of AI conversations creates a direct tension with user privacy. The conversations most important to monitor for safety purposes (expressions of suicidal ideation, disclosures of abuse, mental health crisis communications) are also the conversations users are most likely to want to remain private. A safety monitoring system that stores these conversations creates a database of the most sensitive disclosures that users share with AI systems.

This tension is not merely theoretical. Researchers and journalists have documented cases where mental health AI chat logs have been subpoenaed in legal proceedings, accessed by insiders, or exposed in data breaches. The harm caused by exposure of a user's mental health disclosures can be severe and long-lasting. A safety system that creates this harm while attempting to prevent a different harm is not unambiguously beneficial.

AffixIO's approach resolves the tension by separating what the safety system needs to do (detect distress and trigger a response) from what it does not need to do (store the content that triggered the detection). The distress pattern detector processes content in memory to determine whether a pattern is present. The result is a binary flag: distress_detected = 1 or 0. The flag is passed to the policy enforcement layer. The content is not retained. The safety event record contains the flag, the hashed session identifier, and the timestamp. It does not contain the triggering content.

SECTION 4

Distress Pattern Detection

The distress pattern detector is a keyword and pattern matching component that operates on both the user's query and the AI's response. It checks for patterns associated with three safety domains: self-harm and suicidal ideation, acute crisis and immediate danger, and child safeguarding concerns.

The pattern matching uses a graded escalation structure. Level A patterns (general mental health distress indicators) trigger safety resource inclusion in the AI response without withholding the response. Level B patterns (specific self-harm or suicidal ideation language) trigger the full safety response pipeline: the AI response is withheld, a structured crisis resource response is returned, and a safety event is logged. Level C patterns (immediate danger or CSAM indicators) trigger the Level B response and additionally generate an escalation alert for human review.

TIER	EXAMPLE PATTERNS	RESPONSE	LOGGED
1 (Distress indicators)	Hopelessness, feeling trapped, can't go on	Include crisis resources in response	Safety flag in governance record
2 (Self-harm/suicidal language)	Specific method, active planning language	Withhold response; serve crisis resources	Safety event ZK record + hashed ID
3 (Immediate danger / CSAM)	Immediate harm indicators, illegal content	Level B response + escalation alert	Safety event ZK record + escalation queue

The pattern lists for each tier are maintained by AffixIO's safety team and reviewed quarterly in consultation with mental health professionals and child safeguarding specialists. Pattern lists are not published publicly (to prevent circumvention) but are available to regulatory supervisors under appropriate confidentiality arrangements.

SECTION 5

The Safety Response Pipeline

When a Level B or Level C pattern is detected, the safety response pipeline executes in place of the normal governance pipeline. The AI's response is discarded. The policy enforcement layer prepares a structured safety response containing crisis resource information appropriate to the user's

apparent jurisdiction (UK-specific resources if the service is UK-deployed, generic international resources otherwise). The safety response is returned to the user in place of the AI response.

The safety response is not an AI-generated response. It is a pre-drafted, human-reviewed structured response that includes the appropriate crisis helpline number, a link to immediate help resources, and a brief statement that the conversation topic requires human support. The pre-drafted nature of the response is intentional: an AI-generated response to a crisis situation may be less reliable than a carefully drafted human response.

> **UK crisis resources included in safety responses:** Samaritans (116 123), Crisis text line (text SHOUT to 85258), Childline (0800 1111 for under-18 users), NHS 111 option 2 (mental health crisis), and emergency services (999) for immediate danger. Resource selection is based on pattern tier and detected context.

SECTION 6

Hashed Identifier Logging

When a safety event is logged, the session identifier is hashed before storage. The hashed identifier allows the safety team to identify patterns (for example, repeated safety events from the same session, which may indicate a user in ongoing crisis) without storing a directly identifying session identifier in the safety event record.

The hash is computed as `SHA-256(session material and salt)` where the salt is a randomly generated 32-byte value that is stored separately from the safety event log. The salt-based hash is a one-way transformation: without the salt, the hashed identifier cannot be linked to the session nonce, and therefore cannot be linked to any user identifier derived from the session. The salt is rotated monthly; hashed identifiers computed with the previous salt remain in the log but cannot be linked to current sessions after the rotation.

This approach allows pattern detection (multiple safety events from the same session within a short time window, indicating potential ongoing crisis) without permanent session linkability. It satisfies the GDPR data minimisation requirement while preserving the safety monitoring capability needed for regulatory compliance.

SECTION 7

The Safety Event ZK Record

Each safety event generates a ZK proof record in addition to the hashed identifier log entry. The safety event proof uses the primary policy gate circuit with the topic_signal witness set to 0 (indicating the topic was not safe) as the primary circuit input alongside the other policy conditions. The circuit output is NO, indicating that the response was withheld. The proof is anchored in the same Merkle tree as regular governance events.

The ZK record for a safety event proves that a safety-triggering condition was detected and that the normal AI response was withheld. This is the regulatory evidence that the safety guardrail functioned correctly for this specific interaction. The record does not reveal the nature of the safety concern: an auditor knows that a safety event occurred, not whether it was Level A, 2, or 3, and not what content triggered it.

The separation between the hashed identifier log (which records the tier and hashed session identifier) and the ZK governance record (which proves the circuit outcome) allows different access levels to apply to each. Regulatory supervisors with appropriate authority may access the hashed identifier log under confidentiality arrangements. The ZK governance record is publicly verifiable by design.

SECTION 8

Safe Topic as a Circuit Witness

The `topic_signal` witness in the primary policy gate circuit is set to 0 when a Level B or Level C pattern is detected in the AI response or the user query. When `topic_signal = 0`, the circuit output is 0 (NO) regardless of the other witnesses, because the AND gate requires all inputs to be 1 for a YES output. This means a safety event produces a governance record indistinguishable from other NO outcomes by circuit logic (the circuit output is NO in all cases where any witness is 0).

The indistinguishability is intentional. A governance record that explicitly identified "this NO was due to a safety event" would reveal the nature of the outcome in the public Merkle tree, which is undesirable from a user privacy perspective. The ZK property of the proof means that the reason for NO is not recorded in the publicly verifiable governance record. The safety event log, accessed separately with appropriate authority, provides the additional context needed for regulatory oversight.

SECTION 9

Escalation and Human Review

Level C events generate an escalation alert to the human review queue. The alert contains the hashed session identifier, the timestamp, and the Level C classification. It does not contain the triggering content. Human reviewers can use the hashed session identifier to look up the salt-based session context in the hashed identifier log, which contains the tier and approximate time range of the safety event but not the content.

The human review process for Level C events is time-boxed to 15 minutes from alert generation. The reviewer assesses whether the available context (tier, time, session pattern from the hashed identifier log) indicates an immediate danger situation requiring referral to emergency services or a safeguarding authority. The reviewer does not have access to the conversation content as a matter of system design: the content is not stored anywhere in AffixIO's infrastructure.

This is a deliberate architecture choice. A system that stores conversation content for safety review creates content that can be subpoenaed, breached, or misused. A system that detects and responds at the infrastructure layer without content retention provides the safety response without creating the data retention liability. The trade-off is that human reviewers cannot assess the specific content of a safety event; they can only act on the pattern information available from the hashed identifier log.

SECTION 10

Regulatory Evidence

The safety guardrail system generates three categories of regulatory evidence: the ZK proof records in the Merkle tree (proving that safety evaluations were performed for every interaction), the hashed identifier log (providing aggregate safety event statistics and pattern information), and the quarterly safety pattern review records (demonstrating that the pattern lists are maintained by competent professionals).

Together these three categories satisfy the Online Safety Act's requirement for systems to identify and address safety risks: the ZK records prove the system was active, the hashed identifier log shows what safety events occurred and at what rate, and the review records demonstrate ongoing maintenance. They also satisfy the EU AI Act Article 9 risk management requirement: the safety guardrail is a documented, operational risk management measure with a verifiable audit trail.

The evidence is particularly strong for Online Safety Act purposes because the ZK proof records cannot be altered or back-dated. A regulator examining the Merkle tree cannot be deceived about when safety evaluations occurred or how many took place. The tamper-evidence of the audit trail strengthens the regulatory position substantially compared to a system that provides only administrative records of safety monitoring activities.

SECTION 11

Limitations

The cryptographic distress guardrail has inherent limitations that operators should understand before deployment. Keyword and pattern matching is less accurate than AI-based safety classifiers: it produces both false positives (non-distress content matching safety patterns, triggering unnecessary safety responses) and false negatives (distress content expressed in ways that do not match the pattern library, escaping detection). The false negative rate is the more serious concern in a safety context.

Pattern lists must be maintained and updated. New expressions, slang, and methods emerge continuously. A pattern list that was comprehensive when written may miss novel expressions of the same underlying safety concerns. Quarterly review reduces but does not eliminate this risk. Operators with resources to augment keyword matching with real-time pattern learning should do so.

The system cannot prevent harm that has already occurred by the time the AI interaction begins. It can provide resources and escalate, but it cannot undo harm or guarantee that a user in crisis will use the provided resources. Safety at the infrastructure layer is a necessary but not sufficient component of a comprehensive safety programme.

SECTION 12

Conclusion

The cryptographic distress guardrail demonstrates that AI safety and user privacy are not in fundamental conflict. Safety monitoring at the infrastructure layer using pattern detection, hashed identifier logging, and ZK-anchored safety event records provides Online Safety Act and EU AI Act compliance evidence without creating a database of sensitive conversation content. Users in distress receive an appropriate safety response; the safety event is cryptographically recorded; the conversation content is not retained.

This architecture is more privacy-protective than conventional content-retention safety monitoring, and provides stronger regulatory evidence, because the ZK-anchored safety event records are independently verifiable and tamper-resistant. The combination of infrastructure-layer detection, pre-drafted safety responses, and cryptographic audit records is AffixIO's contribution to the emerging field of privacy-preserving AI safety.

Related reading

- WP-008: Zero-Knowledge Proofs as GDPR Article 25 Infrastructure
- WP-004: Real-Time Zero-Knowledge Governance in the AI Response Pipeline
- WP-003: The Proof-Not-Log Paradigm for AI Audit Trails

Frequently asked questions

Can you detect distress without storing messages?

Pattern tiers trigger at the infrastructure layer; only hashed identifiers and tier classifications enter the audit record.

How does this relate to the Online Safety Act?

Age assurance and harm mitigation duties can be met with privacy-preserving eligibility and safety proofs rather than full content retention.

What happens on Level C events?

Escalation alerts fire with hashed session context for human review within 15 minutes, without exposing message plaintext.

AffixIO | affix-io.com | hello@affix-io.com

All whitepapers | Download PDF

- About
- Solutions
- Legal
- Trust & Security

Contact

truth layer | yes | no | proof