

First Report of Injury — AI Negligence

Intake System Design Specification v0.1

PRE-RELEASE DRAFT

FORENSIC INTAKE DESIGN

WHISTLEBLOWER PROTECTED

Oregon / England & Wales

ARCHITECTURE

§1 System Layers

INTAKE

§2 Form Field Spec

§3 JSON Data Schema

TRIAGE

§4 Regulatory Matrix

§5 Risk Scoring

EVIDENCE & OUTPUT

§6 Chain of Custody

§7 Escalation Bridge

REFERENCE

§8 Legacy FROI Gap

§1 System Architecture — Three-Layer Pipeline

Three sequential layers: submitter-facing intake → forensic triage backend → multi-jurisdiction regulatory output. Each layer produces a cryptographically-signed payload before handoff. No downstream layer can modify an upstream record — a hash mismatch is itself an evidentiary artifact.

LAYER 1 · SUBMITTER-FACING

Dynamic Intake UI — albertlane.net/froi-intake

Model Identification	Harm Vector Selector
Autonomy Gradient (0-100)	Prompt / Output Capture
Evidence Upload + SHA-256	Affected Population Estimate
Jurisdiction Selector	Whistleblower Flag

↓ SHA-256 signed canonical payload

LAYER 2 · FORENSIC BACKEND

Triage & Evidence Engine

Cryptographic Hash Lock	$Risk\ Score\ R = S \times (1-A) \times \log_2(E+1)$
Pattern Frequency Analysis	Append-Only Ledger
Whistleblower Identity Insulation	Kill-Switch Threshold Monitor

↓ severity-gated escalation payload

LAYER 3 · REGULATORY OUTPUT

Multi-Endpoint Escalation Bridge

FTC § 5 Complaint	EEOC Charge Format	CFPB Consumer Complaint
OSHA 300 Supplement	EU AI Act Art. 73 Report	
UK ICO / AISI Notification	Internal Corporate Risk Webhook	

Separation of concerns: A corporate defendant cannot suppress Layer 3 regulatory filings by modifying Layer 1 submission data — the telemetry hash computed at submission time would not match. Evidence tampering is self-documenting.

§2 Intake Form Field Specification

Extended field set for `albertlane.net/froi-intake`. Fields marked NEW extend the current form. Five existing fields (Title, Narrative, Sector, Email, Evidence) are retained and unchanged.

Field	Type	Specification
<code>title</code>	<code>string</code>	Short incident identifier. Max 120 chars. Retained from current form.
<code>narrative</code>	<code>text</code>	Free-form incident description. Min 50 chars encouraged. Retained.
<code>sector</code>	<code>enum</code>	Existing dropdown. Retained. Add: "Agentic / Autonomous System" as a new option.
<code>contact_email</code>	<code>email?</code>	Optional. Retained. Encrypted at rest. Used only for submission receipt and consent-based follow-up.
<code>evidence_file</code>	<code>file?</code>	PNG / JPEG / WebP ≤ 5MB. SHA-256 custody hash issued on upload. Retained.
<code>harm_vector</code> NEW	<code>enum[]</code>	Multi-select, at least one required. Options: discriminatory_bias · hallucinated_negligence · unintended_agent_action · psychological_strain · privacy_breach · financial_harm. Determines regulatory endpoint routing in Layer 3

3.			
model_provider	NEW	string?	Free text. E.g. "OpenAI", "Anthropic", "Meta", "xAI", "Unknown". Routes regulatory payload generator.
model_identifier	NEW	string?	Model name + version as reported by API or UI. E.g. "gpt-4o-2024-08-06". A SHA-256 hash of model_identifier + API response headers is computed server-side for immutability.
autonomy_level	NEW	int 0-100	Required. Slider: 0 = Fully Autonomous (no human in loop) → 100 = Strict Manual Oversight at every step. Inverted in risk formula: low autonomy_level = high autonomy factor A in R. UI label at 0: "No human in loop." UI label at 50: "Human-on-the-loop." UI label at 100: "Human-in-the-loop."
system_prompt	NEW	text?	System / developer prompt active at incident time, if known. Included in telemetry hash. Treated as sensitive: redacted in all public summaries. Used to establish whether harm originates from foundation model or operator configuration.
user_input	NEW	text?	Exact user message or prompt submitted to the system at time of incident. Included in telemetry hash chain.
model_output	NEW	text?	Exact harmful output. The primary evidentiary artifact. Included in telemetry hash. If output was truncated or paginated, submitter should capture the harmful segment

			verbatim.
tool_trace	NEW	json? / file?	For agentic incidents: JSON tool / API calls made by the in sequence. Upload as file d paste raw. Required when harm_vector includes unintended_agent_act Must include: tool name, inp outputs, authorization conte timestamp per call.
affected_count_est	NEW	int?	Estimated number of people affected (E in risk formula). Logarithmic effect on risk score — err toward higher estimate under uncertainty. A single known victim = 1. A pattern affecting hiring pipeline = submitter's estimate of total applicant pool.
incident_datetime	NEW	datetime?	ISO 8601. When the harmful output or action occurred — not when this report is filed. Statute of limitations anchor: preserves claimant's position from incident date, not submission date.
is_recurring	NEW	bool	Has this or substantially similar harm occurred before, from this same system? Triggers pattern-frequency accumulator in the triage engine. Repeated incidents against same model_identifier escalate kill-switch monitoring.
whistleblower_flag	NEW	bool	Submitter is an employee or contractor of the deploying entity. Activates additional identity protection layer. Routes submission to whistleblower-specific legal guidance and evidence lock-in procedures.


```
    "autonomy_level":      8,                                // 0-100 %
    "autonomy_label":      "Human-on-the-loop",
    "system_prompt":       "[REDACTED IN PUBLIC SUMMARY]",
    "temperature":         0.7,
    "rag_chunks_used":      null
  },

  "telemetry": {
    "user_input":           "[exact prompt submitted]",
    "model_output":         "[exact harmful output – primary & secondary]",
    "tool_trace":           null,
    "telemetry_hash":       "sha256:c4e7..."              // SHA-256 hash
  },

  "impact": {
    "affected_count_est":   3400,
    "severity_score":       7,                                // 1-10; 10 = max
    "is_recurring":        true,
    "incident_datetime":    "2026-09-15T09:00:00Z"
  },

  "evidence": {
    "files": [{
      "filename":           "screenshot-2026-09.png",
      "mime_type":          "image/png",
      "file_hash":          "sha256:d4e1...",
      "received_at":        "2026-10-05T14:32:01Z",
      "custody_chain":      ["submitter", "frois-intake-server"]
    }]
  }
}
```

§4 **AI Harm × Regulatory Endpoint Matrix**

Maps each harm vector to the minimum telemetry required for evidentiary sufficiency, and the regulatory endpoints that accept each harm type. A single FROI-AI record can trigger multiple endpoints simultaneously via Layer 3.

Harm Vector	Minimum Telemetry	Primary Endpoints	Secondary Endpoints
Discriminatory / Algorithmic Bias HIGH SEVERITY	model_output, decision log, demographic distribution of outcomes, system_prompt if	EEOC CFPB HUD	State AG State AG

	available			
Hallucinated / Factual Negligence CONTEXT-DEPENDENT	Exact model_output, RAG context chunks at inference time, user_input, model_identifier + version hash	FTC § 5	St Me	
Unintended Agent Action CRITICAL IF FINANCIAL	Full tool_trace (JSON), API call log, autonomy_level ≤ 20, authorization scope in effect	DOL OSHA	SE	
Psychological / Cognitive Strain EMERGING DOCTRINE	Interaction frequency logs, escalation loop documentation, duration of exposure, employer deployment policy	OSHA DOL	St	
Privacy Breach / Data Exfiltration HIGH SEVERITY	Data categories exposed in model_output, RAG retrieval trace, timestamp of first exposure	FTC § 5 GDPR / EU AI Act	UK	
Financial Harm — Credit / Insurance HIGH SEVERITY	Decision output, financial product category, demographic proxies traceable in model config, affected_count_est	CFPB EEOC	St St	

§5 Probabilistic Risk Scoring Algorithm

The triage engine computes a risk score R for each submission. R gates automated escalation — webhooks, kill-switch advisories, and regulatory filings activate only at defined thresholds, preventing noise from drowning out genuine harm signals.

$$R = S \times (1 - A/100) \times \log_2(E + 1)$$

- S** Severity score (1–10). Assigned from harm_vector: physical harm ≥ 9 · discriminatory bias 7–8 · financial loss 6–8 · privacy breach 6–7 · psychological strain 4–6 · reputational 2–4.
- A** autonomy_level field (0–100). Inverted: 0 (fully autonomous) contributes maximum weight; 100 (full human oversight) drives autonomy factor to zero, halving the risk score contribution.
- E** affected_count_est. Logarithmic dampening prevents a single mass-impact figure from dominating without verification. Uncertainty in E: always use higher estimate — conservative rounding is legally defensible.

R < 5

LOG & MONITOR

Record appended to ledger. No automated external action. Internal dashboard flag only. Accumulates toward pattern analysis.

5 ≤ R < 20

INTERNAL ESCALATION

Webhook fires to corporate risk counsel. 72-hour acknowledged response required. Pattern analysis triggered across model_identifier.

R ≥ 20

MANDATORY REGULATORY

Layer 3 regulatory payload generators activate automatically. Kill-switch advisory issued to provider. Legal hold on all related evidence activates.

Kill-switch condition: When 3 or more submissions against the same `model_identifier` each breach $R \geq 20$ within any 30-day rolling window, the system generates an automated notice to the model provider's published safety contact and, where a regulatory sandbox exists, to the relevant governing body. This is the mechanism bridging individual incident to systemic pattern enforcement — the point at which "isolated user error" becomes an untenable corporate defense.

§6 Chain of Custody & Evidence Preservation

The existing site already issues SHA-256 custody receipts per uploaded file — this extended specification covers the full telemetry payload hash.

not just the image file. Each component's inclusion is independently defensible in litigation.

Hashed Component	Why Included	Defensive Purpose
submission_id + submitted_at	Timestamp-anchors the incident before any post-incident model weight changes or log rotations	Defeats: "the model doesn't behave that way anymore"
model_identifier	Links harm to a specific model version, not a product family	Defeats: version confusion — "that was the old model" — because the new one has a different hash
system_prompt (if provided)	Establishes whether operator configuration amplified or caused harm	Allocates liability between foundation model provider and deploying operator
user_input + model_output	Core evidentiary artifact — the actual harm event, verbatim	Immutable. Subsequent model behavior changes are irrelevant to this incident record.
tool_trace (agentic cases)	For autonomous actions: what the agent did, in sequence, with authorization context at each step	Establishes the proximate cause chain — required to hold the deployer liable for consequential agent actions
evidence_files[] hashes	Screenshot / photo custody (existing behavior, retained)	Prevents alteration of visual evidence after submission. Each file hash is part of the outer payload_hash.

Append-only guarantee — all layers: The existing intake stores submissions append-only. The extended system carries this guarantee through all three layers. No record in the ledger may be modified — only superseded by a new submission that references the original `submission_id`. Supersession events are themselves logged and hashed. A deleted-then-resubmitted record leaves a gap in the ledger sequence that is itself evidence of tampering.

§7 Escalation Bridge — Regulatory Payload Formats

When $R \geq 20$ or manual escalation is requested, Layer 3 compiles jurisdiction-specific payloads from the canonical record. Submitters do not re-enter data — each format is derived from the same source fields.

Endpoint	Filing Vehicle	Key FROI-AI Fields
FTC § 5	ReportFraud.ftc.gov complaint schema + narrative export	harm_vector, model_output, model_provider, affected_count_est, narrative
EEOC	OIS pre-charge questionnaire field mapping	harm_vector=discriminatory_bias, model_output, employer/deployer identity, demographic impact
CFPB	Consumer Financial Protection Complaint Database submission	harm_vector=financial_harm, model_provider, narrative, contact_email (optional)
OSHA Supplement	Proposed AI Workplace Harm annex to OSHA 300 log	harm_vector=psychological_strain unintended_agent_action, autonomy_level, employer context
EU AI Act Art. 73	Serious Incident Report — JSON-LD to national market surveillance authority	All telemetry fields, model_identifier, risk_tier, payload_hash, affected_count_est
UK ICO / AISI	AISI incident report form; ICO data breach notification (GDPR Art. 33)	harm_vector=privacy_breach, model_output, data categories, affected_count_est
Corporate Risk Webhook	POST to corporate risk management API; JSON; signed with HMAC- SHA256	Full canonical record + risk_score + triggered_endpoints. $R \geq 30$: adds CISO flag + PR triage signal.

Webhook delivery guarantee: The corporate escalation webhook must require a signed acknowledgment receipt. Unacknowledged webhooks that fail silently are how "we had a process, I promise" defenses get constructed post-litigation. A failed webhook leaves the incident in a

constructed post-incident. A failed webhook leaves the incident in a pending-escalation state visible on the administrative dashboard; it is never silently dropped.

§8 Legacy FROI Gap Analysis
