TRUST FORENSICS, LLC

PRACTITIONER FRAMEWORK · VERSION 1.0

Agent Incident Readiness Framework

A method for testing whether an organization could reconstruct an AI agent incident from its retained records: inventory, evidence map, a 48-hour drill, gap remediation ordered by cost, a 0–100 readiness score mapped to underwriting and governance frameworks, and a first-hour preservation playbook.

Published October 21, 2026 · Last reviewed October 21, 2026 · Version 1.0 · trustforensics.co

1. Purpose and scope

1.1 Purpose

This framework tests one capability: whether an organization can reconstruct what an AI agent did, why it did it, and who authorized it, using records retained before the incident. It covers the inventory, the evidence sources, an exercise that proves the capability on a real action, the usual gaps and their fixes, a score, and the steps for the first hour after an incident.

The framework does not assess model quality, output accuracy or the business case for a deployment. It does not certify compliance with any standard. Section 6 maps its criteria to external frameworks so that results can support, but not replace, those assessments.

1.2 Definitions

- **Agent.** A software system that uses one or more AI models to select and execute actions through tools, APIs or other systems, with some degree of autonomy between human instructions.
- **Harness.** The software that runs the agent: it assembles context, calls the model, interprets tool calls and executes them. Also called the orchestrator or agent framework.
- **Action surface.** The set of actions an agent can execute with the credentials and tools available to it, whether or not those actions were intended.
- **Agent incident.** Any agent action, or chain of actions, that results in or could reasonably result in: loss or corruption of data; an unintended financial transaction; an unintended external communication; an action outside the agent's authorized scope; access to systems or data the agent was not intended to reach; or harm to a third party. Near misses stopped by a control are in scope for logging and review.
- **Reconstruction.** An account of an incident that a third party can verify from retained records. An account that depends on the recollection of staff, or on the agent's own after-the-fact explanation, is not a reconstruction.

1.3 Agents in scope

An agent is in scope if its action surface includes any of the following:

Class	Examples
Write	Creates or modifies records, files, code, configuration, tickets
Delete / destructive	Deletes data, volumes, resources, accounts; overwrites without versioning
Financial	Initiates, approves or modifies payments, refunds, orders, invoices, trades
External communication	Sends email, messages, posts or documents outside the organization
Privilege	Creates, changes or uses credentials, roles or access grants

Read-only agents are out of scope unless they can reach credentials that would give them any of the above. Scope is decided by action surface, not by intended use.

---

2. Agent inventory

Maintain one record per agent in scope. Review the inventory at least quarterly and after any change to the model, harness, tools or credentials.

Field	Content
Agent ID and name	Unique identifier; human-readable name
Business owner	Named individual accountable for the agent's purpose and outcomes
Technical owner	Named individual accountable for its configuration and credentials
Purpose	One-sentence statement of what the agent is for
Model(s)	Provider, model name and version; how the version is pinned or updated
Harness / orchestrator	Product or framework, version, hosting location
Vendors in the chain	Every third party that processes the agent's requests or executes its actions (model provider, harness vendor, observability platform, tool and API providers)
Tools and integrations	Each tool, with the system it acts on
Credentials	Each credential reachable by the agent, its issuer, scope as issued, rotation date, storage location
Action surface	Classes from §1.3 that apply, determined from credentials, not intent
Human approval points	Which actions require human approval; how approval is enforced (in the harness, in the target system, or by instruction only)
Logging	What is logged at each layer, where, and for how long (cross-reference §3)
Kill switch	How the agent is suspended; the named person with authority to do it; the named alternate

Field	Content
Last reviewed	Date and reviewer

Note on credentials. Record every credential the agent could reach, including tokens in configuration files, environment variables, repositories and local machines. The intended credential set and the reachable credential set often differ.

---

3. Evidence map

For each evidence source, record the fields below. Gaps in this table are the main input to the score.

Source	What it contains	Typically controlled by	Retention to confirm	Export method	Chain-of-custody notes
Harness traces	Each step: context assembled, model call, tool call with full request and response payloads, results, errors, timing	Deployer (self-hosted) or harness vendor (hosted)	Default trace retention; whether payloads are stored or only metadata	API or bulk export; screenshot is not an export	Export with hash at collection; record exporter, time, tool and query used
Model provider request/response logs	Prompts and completions as received and returned by the provider; request IDs; timestamps	Model provider	Contracted retention; zero-retention settings; abuse-monitoring retention; whether the customer can retrieve logs at all	Usually by request under contract, if available; confirm the process in advance	Record request IDs in harness traces so provider records can be matched; send preservation request in the first hour
Tool-side and target-system logs	Authoritative record of what was executed: ERP and finance audit trails, bank and payment logs, email server logs, cloud provider API audit logs, database logs	Each system owner or vendor	Native retention per system; tier-dependent limits on SaaS audit logs	Native audit export; cloud audit-log export	These logs are generated outside the agent's control and carry the most weight; preserve first
Observability platform	Aggregated traces, metrics, evaluations, alerts	Observability vendor	Plan-dependent retention; sampling rates	Platform export	Check whether sampling dropped the relevant session

Source	What it contains	Typically controlled by	Retention to confirm	Export method	Chain-of-custody notes
Configuration and prompt version history	System prompts, instructions, rules files, tool definitions, model and parameter settings, harness configuration, with change history	Deployer (repository); harness vendor for its own system rules	Repository history; whether runtime configuration is versioned	Repository export at a commit hash; configuration snapshot	Establish which version was live at the time of the incident; vendor-side system rules may not be visible
Approval and access records	Human approvals and rejections; who approved the deployment and its scope; credential issuance and changes; role assignments	Deployer; identity provider; target systems	Identity provider and approval-tool retention	Native export	Approvals given in chat or email must be collected from those systems
Inputs the agent acted on	Documents, emails, tickets, records, web content that entered the agent's context	Various	Often the shortest: mailboxes, temporary files, fetched web pages	Copy with metadata	Hash each item at collection (SHA-256); record source location and retrieval time

3.1 Integrity principle

Records the agent can write into are statements, not evidence of what happened. This includes its own transcript and its own explanation after the event. Corroborate every material step against at least one record generated outside the agent's write path: a target-system audit log, a provider log, or an infrastructure log.

---

4. The drill

4.1 Design

The drill is a 48-hour reconstruction exercise performed on one real, recent agent action. It tests the records, not the staff.

1. **Select.** The assessor selects an action executed by an in-scope agent between 7 and 30 days earlier. Choose an action with consequence (a write, send, payment or delete), not the most convenient one. The organization is not told which action in advance.
2. **Start the clock.** The organization has 48 hours to produce records answering the five questions in §4.2.

3. **Produce.** Records are exported, hashed and delivered to the assessor. Verbal explanations are noted but do not score.
4. **Verify.** The assessor checks each answer against the records and against at least one independent source.
5. **Report.** Results feed the score (§6) and the gap list (§5).

4.2 Questions and rubric

#	Question	What a complete answer contains
Q1	What did the agent see?	The full context for the decision: instructions, retrieved inputs, prior steps, tool results, with inputs hashed
Q2	What did it decide?	The model output and tool call that produced the action, with full payload
Q3	What authorized it?	The credential used and its scope; the approval record (or the configuration showing none was required); the deployment approval
Q4	What changed since?	Configuration, prompt, model or credential changes between the action and today; whether the system can still be returned to its state at the time
Q5	Can each answer be proven from retained records?	Each answer traced to an exported record, with a corroborating record outside the agent's write path

Score each of Q1–Q4 on this scale:

Level	Description
0	Cannot be answered
1	Answered from recollection or from the agent's own explanation only
2	Partially answered from records; material gaps
3	Fully answered from records, but beyond 48 hours or requiring vendor goodwill
4	Fully answered from retained records within 48 hours

Score Q5 on the same scale, by the weakest of Q1–Q4 after corroboration. The drill result is the sum (0–20).

---

5. Gap categories and fixes

Gaps are listed in ascending order of typical cost and effort. Close the cheaper gaps first: the first four cost a decision or a configuration change.

#	Gap	Fix	Typical effort
G1	No named suspension authority	Name a person, and an alternate, with authority to suspend each agent without further approval; test the kill switch	Governance decision

#	Gap	Fix	Typical effort
G2	Payload logging off, or metadata only	Enable full request/response payload capture in the harness and observability platform for in-scope agents; disable sampling for them	Configuration change
G3	Retention shorter than likely discovery and notice periods	Set retention for traces, tool-side audit logs and approval records to at least 12 months, or to policy notice and limitation periods if longer; export to storage the deployer controls	Configuration; storage cost
G4	Request IDs not linked across layers	Record provider request IDs and target-system transaction IDs in the harness trace	Minor engineering
G5	Configuration and prompts not versioned	Keep all prompts, rules, tool definitions and harness settings in version control; record the live version in each trace	Minor engineering; process
G6	Inputs not preserved	Hash and store inputs that enter context for in-scope actions, or ensure the source system retains them for the retention period	Engineering; storage
G7	Over-scoped or unknown credentials	Inventory reachable credentials; replace broad tokens with least-privilege, purpose-specific credentials; remove credentials from files the agent can read	Engineering; ongoing
G8	No enforced approval for irreversible actions	Require human approval for delete, financial, external-send and privilege actions, enforced by the harness or the target system, not by instruction alone; set value thresholds for financial actions	Engineering; workflow design
G9	Backups reachable by the same credential	Separate backup storage and deletion rights from the agent's credentials; enable delayed or soft delete where available	Infrastructure
G10	No vendor preservation terms	Add to vendor contracts: log retention period, customer right to retrieve logs, preservation on request within a stated time, notification of vendor-side incidents	Negotiation

6. Scoring

6.1 Readiness score (0–100)

The score has ten weighted domains. Each domain is assessed at level 0–4 using the criteria below. Domain points equal weight × level ÷ 4.

Domain	Weight	Level 4 criteria (summary)
D1 Inventory and ownership	10	Every in-scope agent inventoried per §2; owners and alternates named; reviewed within 90 days
D2 Action surface and credentials	12	Reachable credentials inventoried; least privilege; no broad tokens in agent-readable locations
D3 Approval for irreversible actions	12	Enforced approval for all destructive, financial, external-send and privilege actions; thresholds documented
D4 Harness trace logging	12	Full payload traces, unsampled, for in-scope agents; exportable with integrity

Domain	Weight	Level 4 criteria (summary)
D5 Model provider records	8	Provider retention known and contracted; request IDs captured; retrieval process tested
D6 Tool-side audit logs	10	Audit logging on for every target system; export tested
D7 Configuration and prompt versioning	8	All configuration versioned; live version recorded per trace
D8 Input preservation	6	Inputs hashed or retained for in-scope actions
D9 Retention and vendor preservation terms	10	Retention meets §5 G3 across all layers; vendor contracts include preservation terms
D10 Suspension and incident procedure	12	Named suspension authority; kill switch tested; first-hour playbook (§7) adopted; carrier notice requirements recorded

Drill cap. The drill result limits the score:

Drill result (0–20)	Maximum score
16–20	No cap
11–15	79
6–10	59
0–5	39

A Q3 ("what authorized it") score of 0 or 1 caps the score at 59 regardless of the drill total.

6.2 Bands

Score	Band	Meaning
85–100	Ready	Could produce a verifiable reconstruction of an incident within 48 hours
70–84	Substantially ready	Reconstruction possible; specific gaps would slow it or leave a question open
50–69	Partial	Reconstruction would depend on vendor cooperation or leave attribution unresolved
30–49	Limited	Key records missing for at least one layer; reconstruction unlikely to settle cause
0–29	Not ready	Account of an incident would rest on recollection and the agent's own explanation

6.3 Mapping to external frameworks

The table shows where each domain supports evidence for external frameworks. References are to: AIUC-1 as released October 15, 2026 (requirement IDs change between quarterly releases; confirm against the current release); NIST AI RMF 1.0 (NIST AI 100-1) subcategories; and ISO/IEC 42001:2023. ISO/IEC 42001 references are given as "aligned with" because control titles were confirmed from secondary sources, not the licensed text. The final column maps to the documentation categories coverage counsel commonly recommend insureds maintain: AI inventory, harm mapping, governance records, vendor contracts, and incident-response records.

Domain	AIUC-1	NIST AI RMF	ISO/IEC 42001 (aligned with)	Coverage documentation
D1 Inventory and ownership	E004 Assign accountability	GOVERN 1.6; GOVERN 2.1	A.4.2 resource documentation; A.3.2 roles and responsibilities	AI inventory; governance records
D2 Action surface and credentials	A003 Limit AI agent data access; B006 Prevent unauthorized AI agent actions; B007 Enforce user access privileges	MAP 4.2	A.9 use of AI systems	Harm mapping
D3 Approval for irreversible actions	D003 Restrict unsafe tool calls (incl. D003.4 human-approval workflows)	GOVERN 3.2; MAP 3.5	A.9 use of AI systems; A.6.2.6 operation and monitoring	Governance records; harm mapping
D4 Harness trace logging	E015 Log AI system activity (incl. E015.2 execution-chain logging); D003.3 tool call log	MEASURE 2.4; MANAGE 4.1	A.6.2.8 recording of event logs	Incident-response records
D5 Model provider records	E006 Conduct vendor and model due diligence	GOVERN 6.1; MANAGE 3.2	A.10.3 suppliers	Vendor contracts
D6 Tool-side audit logs	E009 Monitor third-party access; E015	MEASURE 2.4	A.6.2.8 recording of event logs	Incident-response records
D7 Configuration and prompt versioning	E004 (approval of changes)	GOVERN 4.2; MANAGE 4.1 (change management)	A.6.2.7 technical documentation	Governance records
D8 Input preservation	E015.3 storage, retention and access	MANAGE 4.3	A.7.5 data provenance	Incident-response records
D9 Retention and vendor preservation	E006; E009; E015.3	GOVERN 6.1; GOVERN 6.2	A.10.2 allocating responsibilities; A.10.3 suppliers	Vendor contracts
D10 Suspension and incident procedure	E001–E003 AI failure plans	MANAGE 2.3; MANAGE 2.4; MANAGE 4.3; GOVERN 4.3	A.8.4 communication of incidents	Incident-response records

Other references. Where an agent is a high-risk AI system under the EU AI Act, D4 and D9 support deployer record-keeping obligations: Article 26(6) requires deployers to keep automatically generated logs under their control for at least six months. Article 12 covers logging capability on the provider side, and Article 73 covers serious-incident reporting. The application dates for high-risk obligations were amended in 2026; confirm the current timetable. D2 and D3 address the risk described in OWASP Top 10 for LLM Applications (2025) as LLM06 Excessive Agency, and in the OWASP Top 10 for Agentic Applications as ASI02 Tool Misuse and ASI03 Identity and Privilege Abuse.

7. Preservation playbook: the first hour

Adopt this page as written or adapt it. Assign each step to a named role before an incident.

Minute 0–10: Contain without destroying

1. Suspend the agent using the kill switch. Do not delete the agent, its workspace, its configuration or its logs.
2. Revoke or rotate the credentials the agent used. Record the time and the person who did it.
3. Record the time the incident was detected, by whom, and how.

Minute 10–25: Freeze retention

4. Disable automatic deletion, rotation and expiry on: harness traces, observability data, target-system audit logs, mailboxes and storage involved.
5. Snapshot affected systems and configuration where possible. Do not restore over the affected system until the snapshot is taken.
6. Record the live versions of prompts, rules, tool definitions, harness configuration and model.

Minute 25–40: Export and hash

7. Export harness traces for the affected agent covering at least 72 hours before the incident through the present.
8. Export target-system audit logs for the same window.
9. Collect the inputs the agent acted on. Compute a SHA-256 hash of each export and each input at collection. Log who collected what, when and how.
10. Store exports in a location the agent and its credentials cannot reach.

Minute 40–55: Preservation requests

11. Send written preservation requests to the model provider, harness vendor, observability vendor and each tool or infrastructure vendor in the chain. Include account identifiers, time window, request IDs if known, and a request to confirm preservation in writing.
12. Preserve internal communications about the agent and the incident.

Minute 55–60: Authority and notice

13. Confirm who holds suspension authority and who will lead the reconstruction. Record decisions from this point in a single incident log.
14. Notify your insurance broker, and the carrier where the policy requires, within the notice window in the policy. Check cyber, technology E&O, management liability and any AI-specific policy. Many policies require notice "as soon as practicable" and some specify a number of days. Missing a window can affect coverage.
15. Consider whether contractual or regulatory notification obligations apply, and to whom. Get legal advice.

Do not, in the first hour: restore from backup over the affected system before snapshot; ask the agent to explain what it did and treat the answer as a record; delete or reconfigure the agent; let vendor support tickets stand in for written preservation requests.

---

8. Document control

Item	Detail
Version	1.0
Published	October 21, 2026
Last reviewed	October 21, 2026
Review cycle	Quarterly, and on any new release of AIUC-1 or revision of the referenced frameworks
Owner	Trust Forensics, LLC

References

- AIUC-1 standard (release of October 15, 2026), Artificial Intelligence Underwriting Company. https://standard.aiuc-1.com/ · changelog https://standard.aiuc-1.com/changelog
- NIST AI 100-1, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, January 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
- NIST AI 600-1, *Generative Artificial Intelligence Profile*, July 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- ISO/IEC 42001:2023, *Information technology — Artificial intelligence — Management system*. https://www.iso.org/standard/81230.html
- Regulation (EU) 2024/1689 (EU AI Act), Articles 12, 26 and 73. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
- OWASP Top 10 for LLM Applications 2025. https://genai.owasp.org/llm-top-10/
- OWASP Top 10 for Agentic Applications. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
- Olshan Frome Wolosky, "How Policyholders Can Press For Narrower AI Exclusions," August 2026. https://www.olshanlaw.com/newsroom/publications/how-policyholders-can-press-for-narrower-ai-exclusions
- Lowenstein Sandler, "AI Risks & Insurance Coverage: What Corporate Policyholders Need to Know," October 2025. https://www.lowenstein.com/media/vbhjxurh/v035-ai-risks-insurance-coverage-what-corporate-policyholders-need-to-know.pdf

This framework is published for use by organizations, brokers and underwriters. It is not legal advice and does not certify compliance with any standard.

Trust Forensics, LLC · 2345 Ashland Avenue, Suite 241, Cincinnati, Ohio 45206 · (513) 212-6596 · matthew@trustforensics.co

Report-based; no criminal, government, or expert-witness work. This document is analysis of the public record and is not legal advice.