



Can You Prove What Your AI Did?

AI Security Investigation & Auditability Checklist

AI systems are gaining access to enterprise information, tools, applications and infrastructure. This checklist helps security and governance teams test whether AI-initiated actions can be reconstructed during an investigation or audit — across 40 checks in eight areas.



THE FINAL TEST

If an AI-driven security decision was challenged 90 days from now, could you reconstruct why it was made, what evidence supported it and what happened next?

WHY THIS CHECKLIST

If AI becomes part of the decision chain, it must also become part of the evidence chain.

AI is moving from assisting security teams to acting on their behalf. Agents now summarise alerts, gather evidence, call APIs, change configurations and trigger remediation. Gartner expects that by 2028, 15% of day-to-day work decisions will be made autonomously through agentic AI.¹

That is a real opportunity — and the organisations that capture it fastest will be the ones that can show their boards, auditors and regulators exactly what their AI systems did. Auditability is what allows AI to be given more authority with confidence.

13%

of organisations reported a breach involving an AI model or application

IBM, Cost of a Data Breach 2025²

97%

of those breached organisations lacked proper AI access controls

IBM, Cost of a Data Breach 2025²

63%

of breached organisations had no AI governance policy or were still developing one

IBM, Cost of a Data Breach 2025²

The core question. Can we reconstruct the actions of an AI system as reliably as we can reconstruct the actions of a privileged administrator? AI agents increasingly carry the same characteristics — an identity, credentials, access to sensitive data, connections into enterprise applications and the authority to act. They need the same evidence trail.

The minimum evidence record for an AI-initiated action

Attribute	What to capture
Identity	The agent's own workload or service identity <i>and</i> the human or process it acted on behalf of (the delegation chain)
Timestamp	Synchronised, UTC-normalised time for each step — request, tool call, decision, action
Source	The AI system, model and version, and the orchestrator or interface that invoked it
Target	The system, dataset, API or resource acted upon
Action	The query, tool invocation, configuration change or command performed
Outcome	Success, failure or partial completion, and the resulting state change
Context	Evidence relied on, permissions in force, approval state, and a correlation ID linking every step

How to use this checklist

- **Work through it with the right people** — security operations, identity, the AI platform owner and governance or risk.
- **Score each check on evidence, not intent.** A control counts only if you can show the record.
- **Start with your highest-authority agent** — the one that can change, disable or remediate something.

Yes — 2

In place, and you can produce the evidence today.

Partial — 1

In place for some systems, or not yet tested.

No — 0

Not in place, or unknown.

THE CHECKLIST

Eight areas. Forty checks.

1

AI inventory

Do you know which AI systems and agents are operating in the environment?

SCORE / 10

Why it matters:

you can't log, govern or investigate an agent you don't know is running — including AI features embedded in SaaS platforms and security tools.

CHECK	EVIDENCE TO LOOK FOR	Y	P	N
1.1 A register exists of AI systems and agents in use, including embedded AI features in SaaS and security platforms.	Current AI register or CMDB category, with a named owner per entry	☐	☐	☐
1.2 Each entry records purpose, model or provider, and the systems, data and tools it connects to.	Integration map; connector and API lists per agent	☐	☐	☐
1.3 Each agent's authority level is defined: read-only, recommend, act with approval, or act autonomously.	Documented authority level in the register; matching approval workflow	☐	☐	☐
1.4 There is a discovery process for unsanctioned AI use and new agents.	Periodic review of OAuth app consents, API keys and service principals; outbound traffic to AI provider endpoints	☐	☐	☐
1.5 The register is reviewed on a defined cadence and whenever an agent is deployed or its scope changes.	Review records; change-management tickets referencing the register	☐	☐	☐

2

Identity

Can every AI-initiated action be associated with an identifiable workload, service or agent identity?

SCORE
/ 10

Why it matters: if agents share accounts with people or other automation, every downstream event becomes ambiguous. OWASP lists *Identity Spoofing & Impersonation* (T9) among its agentic AI threats.³

CHECK	EVIDENCE TO LOOK FOR	Y	P	N
2.1 Each agent has its own distinct workload or service identity — no shared service accounts and no use of human accounts.	One identity per agent in the directory or cloud IAM	☐	☐	☐
2.2 Agent identities are clearly distinguishable from human administrators and other automation in logs.	Naming convention, tags or attributes visible in event records	☐	☐	☐
2.3 Agent credentials are managed and short-lived rather than static keys embedded in code or configuration.	Managed identities, secrets vault, rotation records	☐	☐	☐
2.4 Authentication events for agent identities are collected and retained.	Windows 4624 , 4768 / 4769 , 4776 for on-premises accounts; identity-provider service-principal sign-in logs; cloud role-assumption events	☐	☐	☐
2.5 The delegation chain is recorded: which human or process the agent was acting on behalf of.	User or requester ID carried in the agent's request and action logs	☐	☐	☐

3

Permissions

Can you establish what the AI was authorised to access at the time?

SCORE
/ 10

Why it matters: an investigation asks what the agent *could* do on a past date, not what it can do today. OWASP lists *Privilege Compromise* (T3) as an agentic AI threat.³

CHECK	EVIDENCE TO LOOK FOR	Y	P	N
3.1 Each agent's permissions are documented and scoped to its defined authority level (least privilege).	Role and scope definitions per agent identity	☐	☐	☐
3.2 You can reconstruct an agent's permissions at a specific point in the past.	Retained group-membership changes (4728 , 4732 , 4756), object permission changes (4670), directory audit logs and cloud IAM policy and role-assignment events	☐	☐	☐
3.3 Privileged sessions by agent identities are visible.	4672 special privileges assigned at logon; privileged role activation events	☐	☐	☐
3.4 Permission changes made <i>to</i> or <i>by</i> agent identities trigger review or alerting.	Detection rules or review queue covering agent identities	☐	☐	☐
3.5 Periodic access reviews include non-human and agent identities.	Access-review records listing agent identities and outcomes	☐	☐	☐

4

Interaction logging

Are prompts, requests, system interactions and relevant contextual events recorded appropriately?

SCORE

/ 10

Why it matters: the interaction record explains what the AI was asked and what it returned. ISO/IEC 42001 control A.6.2.8 requires organisations to determine when AI event logging is enabled — at a minimum, while the system is in use.⁴

CHECK	EVIDENCE TO LOOK FOR	Y	P	N
4.1 Requests and responses are logged with requester identity, timestamp and a session or correlation ID.	Sample interaction log with all four fields populated	☐	☐	☐
4.2 The model, version and relevant configuration in use are recorded for each interaction.	Model and version fields in logs; configuration change history	☐	☐	☐
4.3 Interaction logs are exported to a central log store, not held only in the AI application's history.	Log forwarding or API export configured; events visible centrally	☐	☐	☐
4.4 A policy defines how sensitive data in prompts and responses is handled in logs.	Masking or redaction rules; access controls on interaction logs	☐	☐	☐
4.5 Logging is enabled across the AI system's operational life and can't be silently switched off.	Logging configuration under change control; alerts on logging disabled	☐	☐	☐

5

Data and tool access

Can you identify which systems, APIs, datasets or tools the AI accessed?

SCORE

/ 10

Why it matters: a finding is only defensible if you can trace it to the evidence it was built from — and show what else the agent could see.

CHECK	EVIDENCE TO LOOK FOR	Y	P	N
5.1 Every tool or API invocation is recorded: tool, target, parameters and result.	Orchestrator or agent-framework tool-call logs	☐	☐	☐
5.2 Data sources queried by the AI are recorded, including scope of the query.	Query logs naming the dataset, index or time range searched	☐	☐	☐
5.3 AI findings can be traced back to the specific records or documents they relied on.	Citations or record references in AI output; retrieval logs	☐	☐	☐
5.4 Target-side logs are collected as well as the agent's own record.	SaaS audit logs, cloud data-access logs and database audit logs for systems the agent touches	☐	☐	☐
5.5 A correlation ID is propagated from the agent into downstream calls where the platform supports it.	Same ID visible across agent, API gateway and target logs	☐	☐	☐

6

Action logging

Are changes, commands and remediation actions recorded independently from the AI application's own history?

SCORE

/ 10

Why it matters: if the system that took an action is the only witness to it, the evidence isn't independent. OWASP's *Repudiation & Untraceability* (T8) describes agent actions that "cannot be reliably traced back or accounted for due to insufficient logging or transparency."³

CHECK	EVIDENCE TO LOOK FOR	Y	P	N
6.1 Actions on endpoints and servers are captured by the target system's own logs.	4688 process creation with command-line auditing; 7045 service installed; 4698 scheduled task created	☐	☐	☐
6.2 Identity remediation actions are captured by the directory.	4725 account disabled; 4724 password reset attempt; identity-provider audit events for session revocation	☐	☐	☐
6.3 Cloud and SaaS changes made by agents are captured in control-plane and admin audit logs.	Cloud management-plane events and SaaS admin audit logs attributed to the agent identity	☐	☐	☐
6.4 Both the <i>recommended</i> and the <i>executed</i> action are recorded and can be linked.	Recommendation record and execution record sharing a correlation ID	☐	☐	☐
6.5 Action logs are protected from modification, including by the agent's own identities.	Logs forwarded off-host promptly; write-once or tamper-evident storage; alerting on 1102 (Security log cleared)	☐	☐	☐

7

Human intervention

Can you identify when a human approved, changed, rejected or overrode an AI-generated action?

SCORE

/ 10

Why it matters: human oversight is only a control if it leaves a record. OWASP also flags *Overwhelming Human in the Loop* (T10) — where approval volume turns review into rubber-stamping.³

CHECK	EVIDENCE TO LOOK FOR	Y	P	N
7.1 The action classes that require human approval are defined for each agent.	Approval policy mapped to the authority levels in 1.3	☐	☐	☐
7.2 Approve, modify, reject and override decisions are logged with the approver's identity and time.	Approval workflow logs or SOAR case history	☐	☐	☐
7.3 Each approval record links to the specific AI recommendation it relates to.	Shared correlation or case ID across recommendation and approval	☐	☐	☐
7.4 Use of break-glass, manual override or agent stop controls is logged and reviewed.	Override events and post-use review records	☐	☐	☐
7.5 Approval volume and time-to-approve are monitored to detect review fatigue.	Approval metrics per reviewer and per agent	☐	☐	☐

8

Retention and investigation

Will sufficient evidence still exist if the decision is questioned weeks or months later?

SCORE

/ 10

Why it matters: questions about AI decisions often arrive long after the event. Under the EU AI Act, logs automatically generated by high-risk AI systems must be kept for at least six months by providers (Article 19) and by deployers, to the extent the logs are under their control (Article 26(6)).⁵

CHECK	EVIDENCE TO LOOK FOR	Y	P	N
8.1 Retention periods for AI interaction, tool and action logs are defined against regulatory and investigation needs.	Retention policy naming AI log types explicitly	☐	☐	☐
8.2 AI evidence is held in the same searchable evidence store as human and system activity.	One query can return agent, user and system events for the same timeframe	☐	☐	☐
8.3 Timestamps are normalised and clock sources synchronised across AI, identity, endpoint and cloud logs.	UTC normalisation at collection; time-sync monitoring	☐	☐	☐
8.4 Retained evidence is integrity-protected and access to it is restricted and logged.	Tamper-evident or immutable storage; access logs for the evidence store	☐	☐	☐
8.5 AI-initiated actions are covered in incident response playbooks, and reconstruction has been tested.	Playbook section on AI actions; results of the 90-day test. NIST SP 800-61 Rev. 3 positions this preparation as part of incident response. ⁶	☐	☐	☐

Windows event IDs refer to the Windows Security log (except 7045, System log) and require the relevant advanced audit policy subcategories to be enabled. This checklist is a self-assessment aid and does not constitute legal or compliance advice.

YOUR RESULTS

Score summary

Area	Key question	Score
1. AI inventory	Do we know which AI systems and agents are operating?	/ 10
2. Identity	Is every AI action attributable to a distinct identity?	/ 10
3. Permissions	Can we show what the AI was authorised to do at the time?	/ 10
4. Interaction logging	Are requests, responses and context recorded centrally?	/ 10
5. Data and tool access	Can we trace what the AI accessed and relied on?	/ 10
6. Action logging	Are actions recorded independently of the AI application?	/ 10
7. Human intervention	Is human approval, change or override on record?	/ 10
8. Retention and investigation	Will the evidence still exist and be usable months later?	/ 10
Total		/ 80

64–80

Evidence-ready

AI actions can be reconstructed with confidence. Focus on testing, and on extending the same controls to each new agent before it goes live.

40–63

Partially reconstructable

Some AI actions can be explained, others can't. Prioritise any area scoring 4 or below — one weak area can break the whole timeline.

0–39

Evidence gap

AI decisions would be difficult to defend if challenged. Start with identity (2), action logging (6) and retention (8) for your highest-authority agent.

Read the lowest area first, not the total. An investigation follows a chain: identity → permissions → data accessed → recommendation → approval → action → outcome. A strong total with one weak link still leaves a gap at that link.

PUT IT TO WORK

The 90-day test

Choose one AI-driven security action from roughly 90 days ago — ideally one that changed something, such as disabling an account, isolating a host or modifying a policy. Then try to answer each question below from retained evidence alone, and note where the answer came from.

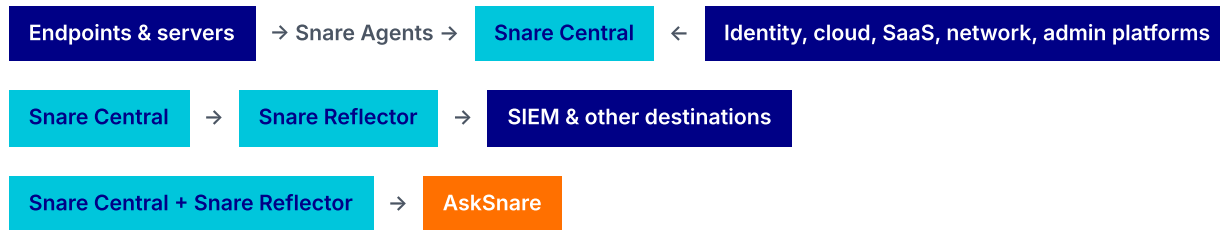
Question	Answer	Evidence source
Which agent identity acted, and on whose behalf?		
What was it authorised to access at that moment?		
What evidence did it rely on to reach its recommendation?		
Which tools, APIs or systems did it invoke?		
What did it recommend, and what was actually executed?		
Who approved, changed or rejected it — and when?		
What happened in the systems affected immediately afterwards?		

Wherever an answer is blank or relies on memory rather than a record, you've found your AI evidence gap.

BUILDING THE EVIDENCE LAYER

Where Snare fits

Most of the checks in this list depend on one thing: a reliable, independent record of activity across the systems AI agents touch. That's the evidence layer — and it's what Snare provides, alongside the SIEM and security tools you already use.



- **Collect from where agents act.** Snare Agents collect events from Windows, Linux and other endpoints and servers. Identity, cloud, SaaS, network and admin-platform sources feed Snare Central directly via syslog, API and cloud audit feeds — no agent required.
- **Keep a complete, independent record.** Snare Central retains a full copy of that evidence for search, reporting and forensic investigation, independently of any AI application or analytics platform (checks 4.3, 6.5, 8.2, 8.4).
- **Route what each destination needs.** Snare Reflector routes selected events to the SIEM and other destinations based on your own rules. Filtering shapes what each destination receives, not what is kept.
- **Investigate faster, from the evidence.** AskSnare, Snare's cybersecurity AI assistant, works from the evidence layer — accelerating investigation while every finding stays grounded in records an analyst can verify.

AI can accelerate an investigation. It shouldn't replace the underlying evidence.

Related resources from Snare

- **Can You Prove What Your AI Did?** — the companion article on AI auditability and the evidence chain.
- **Cybersecurity Visibility Readiness Checklist** — assess collection, coverage, integrity, retention, routing, storage, investigation and recovery.
- **The Intelligent Log Management Playbook** — what to collect, what to retain, what to ingest into your SIEM, and how to keep forensic visibility.

References

1. Gartner, press release, 25 June 2025 (agentic AI predictions: 15% of day-to-day work decisions made autonomously by 2028). Reported in SD Times: sdtimes.com/agentic-ai/gartner-more-than-40-of-agentic-ai-projects-will-be-canceled-in-next-few-years/
2. IBM Security, *Cost of a Data Breach Report 2025*. Reported in Help Net Security, 4 August 2025: helpnetsecurity.com/2025/08/04/ibm-cost-data-breach-report-2025/
3. OWASP GenAI Security Project, *Agentic AI — Threats and Mitigations*: T3 Privilege Compromise, T8 Repudiation & Untraceability, T9 Identity Spoofing & Impersonation, T10 Overwhelming Human in the Loop.
4. ISO/IEC 42001:2023, *Information technology — Artificial intelligence — Management system*, Annex A, control A.6.2.8 "AI system recording of event logs."
5. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Articles 12, 19 and 26(6): ai-act-service-desk.ec.europa.eu/en/ai-act/article-19-and/article-26
6. NIST SP 800-61 Rev. 3, *Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile*, April 2025: csrc.nist.gov/pubs/sp/800/61/r3/final