WP-034
June 2026
12 sections

ENTERPRISE AI GOVERNANCE & COMPLIANCE

Shadow AI Governance: Why Policies Fail and How Cryptographic Proof Fixes It

67% of employees use AI tools their employer has not sanctioned. Only 13% of organisations combine policies with audits. The EU Product Liability Directive arrives in December 2026. Policy is intent. Proof is governance.

AffixIO Research | June 2026 | Download PDF

EXECUTIVE SUMMARY

Shadow AI — the use of artificial intelligence tools without organisational approval or oversight — is the fastest-growing enterprise risk of 2026. Shadow AI-related data loss has increased nearly four times year-on-year. 97% of organisations that experienced AI-related breaches lacked proper access controls. The EU Product Liability Directive, taking effect on 9 December 2026, classifies AI systems as products subject to no-fault liability, making the gap between policy and proof a direct legal exposure. This paper explains why traditional shadow AI governance approaches fail: they detect after the fact, rely on writable logs that auditors will discount, and confuse the existence

of a policy with evidence of compliance. It then presents the technical and architectural requirements for a genuine shadow AI audit trail: runtime evidence of authorised use at the moment of each AI interaction, tamper-evident and cryptographically bound, compatible with GDPR, the EU AI Act, DORA, and HIPAA retention requirements. AffixIO's approach uses Merkle-anchored, ML-DSA-65-signed audit records with zero-knowledge proofs of tool authorisation, enabling organisations to demonstrate compliance without disclosing the content of AI interactions.

CONTENTS

SECTION 01

The Shadow AI Problem You Already Have

In May 2023, Samsung engineers entered proprietary semiconductor source code into ChatGPT to help debug a test sequence. The code left Samsung's network, was processed by OpenAI's servers, and potentially became part of

training data. Samsung discovered the incident after the fact. The employees were not rogue; they were trying to do their jobs faster with a tool that worked. They had no idea the data was leaving the building.

That incident made headlines because Samsung is a publicly identifiable company. The same thing happened, and continues to happen, at most large organisations every day. The tools are different now: GPT-4o, Claude, Gemini, Copilot, Cursor, Perplexity, DeepSeek, dozens of AI-enabled SaaS features your employees didn't have to install because they arrived embedded in tools they were already authorised to use. The problem is no longer whether your employees are using unauthorised AI. They are. The question is whether your organisation can prove what happened, to whom, with what data, and whether it was inside sanctioned boundaries.

The answer, for the overwhelming majority of organisations, is no.

67%

of employees use AI tools at work. Only 18% of companies have AI security policies in place. *(Red Team Partner, 2026)*

4×

increase in shadow AI-related data loss year-on-year. 65% of incidents expose personal data. *(IBM / TechnologyRadius, 2026)*

97%

of organisations that experienced AI-related security incidents lacked proper access controls. *(Multiple industry sources, 2026)*

13%

of organisations combine written AI policies with regular audits for unauthorised AI use. The other 87% have policies without enforcement. *(ISACA / IBM, 2026)*

Shadow AI is not a niche technical problem. It is the gap between the AI tools your organisation approved and the AI tools your employees are actually using, applied to your most sensitive data, every working day. In 2026, it has acquired a hard regulatory deadline: the EU Product Liability Directive takes effect on 9 December 2026 and classifies AI systems as products subject to

strict, no-fault liability. Organisations that cannot document that the AI tools used in a given decision were authorised and compliant at the point of use will face direct legal exposure, not just reputational risk.

This paper explains what shadow AI governance requires, why most current approaches fall short, and what a verifiable audit trail actually looks like.

SECTION 02

Three Forms of Shadow AI in 2026

Shadow AI has evolved beyond the early-2023 pattern of employees copying data into a web chat interface. Understanding what it looks like now is necessary before any governance architecture can address it.

Form 1: Shadow SaaS AI

The original and still largest category. Employees use AI tools directly: ChatGPT, Claude.ai, Gemini, Perplexity, AI writing assistants, AI-powered code review tools, AI legal research platforms. These are not installed on corporate devices; they run in browsers on corporate networks, on personal devices on home networks, or in private browser sessions that bypass corporate web filters. By 2026, Gartner estimates that 70% of AI interactions will occur through features embedded in existing, sanctioned SaaS applications, making it increasingly impossible for IT to distinguish between approved usage of an approved tool and unapproved usage of an unapproved AI feature within that same tool.

Form 2: Shadow Agentic AI

The rapidly growing category. An employee or a small team deploys an AI agent, connecting it to a company email account, a CRM system, a file storage service, or a code repository, via API or a no-code integration platform. The agent runs autonomously, reading and writing data, sending communications, making API calls, and making decisions, without any formal security review. According to Gravitee's State of AI Agent Security 2026 report, 80.9% of technical teams have already moved beyond planning AI agents into active

testing or production deployments, but only 14.4% of those agents went live with full security and IT approval. 47.1% of organisations' AI agents are actively monitored; the rest operate without consistent oversight or logging.

Form 3: Embedded Model Shadow AI

The least visible category. AI capabilities are embedded by software vendors directly into tools that employees already have permission to use. Microsoft 365 Copilot, Salesforce Einstein, GitHub Copilot, Google Workspace AI, Adobe Firefly: all are AI features that arrive inside sanctioned products, process sensitive data, and generate outputs that influence decisions, all without appearing on any shadow AI detection scan because the host application is approved. Of the 2,400 popular business software providers advertising AI capabilities in 2026, 63.6% did not disclose their third-party AI subprocessors in their legal documentation. Employees using these tools for sensitive work are using shadow AI even if they never opened a consumer AI website.

FORM	EXAMPLES	VISIBILITY TO IT	DATA EXPOSURE RISK	GOVERNANCE CHALLENGE
Shadow SaaS AI	ChatGPT, Claude.ai, consumer AI tools	Low (browser-based, BYOD)	High (direct data input)	Detection after the fact; can't prove what was sent
Shadow Agentic AI	Custom agents, no-code automations, API integrations	Very low (no installation footprint)	Very high (autonomous read/write)	No audit trail; agent actions unattributed
Embedded Model AI	Copilot in Word, Salesforce Einstein, GitHub Copilot	Effectively zero (inside sanctioned apps)	Medium-high (processes existing data)	Vendor AI subprocessors undisclosed; no consent obtained

SECTION 03

Why Shadow AI Policies Fail

Most organisations' response to shadow AI has been policy-first: write an acceptable-use policy, communicate it via email and training, and assume that awareness produces compliance. The data shows this does not work. There are five structural reasons why.

Reason 1: Policies Cannot Be Enforced at the Speed of AI Adoption

AI tools launch daily. By the time a new AI product has been reviewed by the security team, approved by legal, communicated in a policy update, and acknowledged by employees, the next generation of the tool has launched. The Samsung incident happened in the very first weeks of ChatGPT's enterprise availability. The employees were not violating a known policy; the policy did not yet exist for the specific tool they used. Policy-based governance operates on quarterly cycles; AI adoption operates on daily ones.

Reason 2: Detection Happens After Exposure

CASB (Cloud Access Security Broker) tools, DLP (Data Loss Prevention) systems, and network monitoring can detect some shadow AI usage after data has already left the organisation. They cannot prevent the exposure; they can only alert security teams that it happened. The Samsung breach was discovered because employees discussed it internally; not because a technical control caught it. Approximately 87% of organisations may lack mature shadow AI detection capabilities, meaning most incidents are not detected at all.

Reason 3: Writable Logs Are Not Audit Evidence

Many organisations believe they have addressed shadow AI governance because their SIEM, their browser proxy logs, or their endpoint agent records AI tool usage. ISACA's 2026 guidance is direct on this point: "Logs stored in writable databases without cryptographic integrity proofs fail modern audit requirements. Auditors trained to spot AI-manipulated evidence will discount logs that cannot prove they have not been altered." A log entry in a mutable

database proves that someone had write access to that database at the time the entry was created. It does not prove that the entry accurately reflects what actually happened. In a regulatory investigation or a PLD liability case, a writable log is not evidence; it is a starting point for questioning.

Reason 4: Policies Don't Produce Runtime Evidence

ISACA's four pillars of genuine AI audit evidence are: (1) who initiated the AI query; (2) what data was accessed and whether that access was authorised; (3) what the active policy and safeguard state was at the exact moment of execution; and (4) which specific model version and data snapshot produced the result. A policy written months before the interaction can satisfy none of these requirements at the moment of runtime. What is needed is not documentation of intended behaviour, but evidence of actual behaviour at execution time.

Reason 5: The Liability Gap in AI Contracts

Even where organisations have formal agreements with AI vendors, those agreements typically provide minimal protection for shadow AI incidents. As Clifford Chance noted in February 2026, standard AI vendor contracts disclaim liability for agent errors, exclude "loss of data" and "consequential damages" from indemnification, and cap liability at subscription fee levels. An organisation whose employee used a shadow AI tool to process customer data, resulting in a GDPR breach, will find that neither the AI vendor's contract nor their internal policy provides a defence against a regulatory penalty or a civil claim. Only a compliance record demonstrating that the AI tools used were authorised and operated within policy boundaries provides that defence.

SECTION 04

The December 2026 Liability Cliff

Shadow AI governance has been a best-practice recommendation since 2023. On 9 December 2026, it becomes a hard legal requirement for organisations operating in or selling into the EU, as the EU Product Liability

Directive (PLD 2022/0302) takes effect in national law across EU member states.

What the PLD Changes

The original 1985 Product Liability Directive did not cover software. The 2022 revision explicitly includes software, digital services, and AI systems within the definition of "products." The critical change is the liability standard: the PLD imposes **strict (no-fault) liability**. This means that if an AI system causes damage, the organisation that placed it on the market or deployed it is liable for that damage, even if they were not negligent.

The PLD also shifts the burden of proof in ways that directly affect shadow AI governance:

- Claimants can request courts to order organisations to **disclose relevant evidence** about an AI system's operation. If you cannot produce evidence of how an AI tool was deployed and what it did, the court may presume defectiveness.
- The PLD links defectiveness directly to **compliance with EU and national product safety law**, including the EU AI Act. A shadow AI tool that was not properly risk-assessed is, by definition, non-compliant with the EU AI Act's Article 9 risk management requirements. Under the PLD, that non-compliance makes it defective. Defectiveness means liability.
- Products placed on the market or put into service **after 9 December 2026** fall within scope. Any AI feature rolled out to employees after that date, without documentation of authorisation and compliance, is a potential PLD liability event from day one.

The Personal Liability Dimension

Alongside the PLD, CISO personal liability for AI governance failures has become a concrete concern: 78% of CISOs reported worrying about personal liability in 2026, up from 56% the previous year. The SEC enforcement actions in the US against individual executives for cybersecurity misrepresentation have established a precedent; EU regulators following the PLD framework will

have similar tools. An executive who signed off on an AI deployment programme without requiring a documented audit trail faces personal exposure if that programme subsequently causes harm.

> **The December 2026 checklist:** For every AI tool in use in your organisation after 9 December 2026, you need to be able to answer: Was this tool authorised? Who authorised it? What data could it access? What were the active controls at the moment of each interaction? Can you prove the answers to these questions from an immutable record that cannot have been altered after the fact?

SECTION 05

Policy Is Intent. Proof Is Governance.

This phrase, from ISACA's 2026 AI audit guidance, captures the entire problem with current shadow AI governance practice. It is worth dwelling on precisely what it means.

A policy says what you intended to happen. It says which AI tools employees were permitted to use, what data classifications those tools were approved for, and what the consequences of non-compliance would be. A policy is a statement of intent at the time it was written. It cannot tell a regulator, an auditor, or a court what actually happened at 14:32 on a given Tuesday when an employee opened a consumer AI tool and pasted in a customer record to help draft a response.

Proof says what actually happened. It says that at 14:32, AI Tool X was invoked, that the interaction was attributed to Employee Y, that the data classification at the time of access was Z, that the active safeguards were Q and R, and that the specific version of the model used was V. It says this in a form that cannot be altered after the fact, that can be verified by any authorised party, and that will be accepted as evidence in a regulatory or legal proceeding.

The gap between policy and proof is not a matter of effort or intent on the part of security teams. It is an architectural gap. Policies are documents. Proof requires an instrumented, tamper-evident record of runtime behaviour. You cannot generate proof by writing a better policy. You can only generate proof by building the infrastructure that captures and preserves evidence at the moment AI interactions occur.

The Four Pillars ISACA Requires

ISACA's 2026 guidance identifies four things that a genuine AI audit trail must document for every AI interaction:

1. **Request origination:** Who initiated the query? Was the initiating identity authenticated? Is the authentication record linked to the AI interaction record?
2. **Data lineage and authorisation:** What data was accessed? Was the data classification checked at access time? Is there a record that the access was within authorised boundaries, not just that authorisation policy existed?
3. **Active controls at execution time:** What safeguards were active at the exact moment the AI system ran? Not what safeguards were supposed to be active; what were actually active, verified at runtime?
4. **Model configuration and data snapshot:** Which specific version of the AI model produced the output? What dataset or knowledge base was it operating on? Was the model the authorised version?

Recording the prompt and response is not an audit trail. It is a transcript. A transcript tells you what was said. It does not tell you whether the access was authorised, whether the controls were functioning, or whether the model version was the one that had been reviewed and approved. Shadow AI governance requires the latter, not just the former.

SECTION 06

What a Valid AI Audit Trail Actually Requires

Given the four ISACA pillars and the PLD evidentiary standard, a valid shadow AI audit trail has the following properties. Each is a technical requirement, not a policy preference.

Immutability

The audit record must be impossible to alter after the fact. A record in a SQL database with write access granted to the application layer, the DBA team, and the infrastructure team is not immutable. A cryptographic hash of the interaction record, embedded in a Merkle tree and signed with an ML-DSA-65 (post-quantum) key, is immutable: altering the record changes its hash, which invalidates the Merkle inclusion proof, which invalidates the signed root. Any alteration is detectable. This is not a configuration choice; it is a mathematical property of the data structure.

Completeness

The audit record must capture all interactions, not a sample. Shadow AI governance that records 10% of AI interactions (because a proxy server drops some traffic) or that only records interactions that pass through a corporate gateway (missing BYOD and home-working usage) provides an incomplete picture that an auditor or court will note. Completeness is achieved by instrumentation at the point of AI tool integration, not by network monitoring.

Attribution

Every interaction record must be attributed to an authenticated identity. Anonymous interaction records ("some user accessed AI Tool X") are not governance evidence. Attribution requires that authentication state at the time of the interaction is captured and bound to the interaction record, not held in a separate system that may be queried independently.

Content Binding Without Content Disclosure

A record must bind the audit evidence to the content of the interaction without necessarily storing the content. This is the zero-knowledge requirement: an auditor needs to be able to confirm that a given AI interaction used an authorised tool, accessed authorised data, and operated under active controls, without the auditor needing to read the content of the employee's query or the AI's response. The content of many AI interactions is itself sensitive: legal advice, HR discussions, medical information, intellectual property. An audit trail that requires storing full interaction transcripts creates a second data liability. Cryptographic commitments to interaction content (hash-bound records) provide evidential binding without content disclosure.

Regulatory Retention Compliance

Records must be retainable for the applicable regulatory period: 10 years for EU AI Act Article 12 (high-risk AI), 5 years for DORA (ICT records), 6 years for HIPAA (security documentation), 5-7 years for MiFID II (financial services). Retention at this scale with per-record signatures would be unmanageable; Merkle-anchored batch attestation (one ML-DSA-65 signed root per batch, individual records verified via O(log n) inclusion proofs) reduces the signature storage requirement by over 99% while maintaining full individual record verifiability.

SECTION 07

Cryptographic Proof of Tool Authorisation

The technical architecture that satisfies all five requirements in Section 6 is a combination of cryptographic commitments, zero-knowledge proofs, and post-quantum anchoring. This section explains how each element contributes.

The Tool Authorisation Registry

Authorised AI tools are recorded in a cryptographically signed registry. Each entry records: the tool identifier (model name, version, API endpoint), the authorisation date, the data classification levels the tool is approved for, the active safeguard configuration, and the authorising officer. The registry is

signed with ML-DSA-65, producing a tamper-evident record of the authorisation state at each point in time. When a tool version is updated or a new tool is added, a new signed entry is created; the history is append-only and immutable.

Runtime Authorisation Check and Interaction Record

At the moment of each AI interaction, the instrumentation layer records:

- A hash of the interaction content (prompt + response), not the content itself
- The tool identifier and version in use
- A reference to the tool's entry in the authorisation registry (proving the tool was authorised at this version)
- The authenticated user identity
- The data classification of the context in which the interaction occurred
- A timestamp
- A cryptographic commitment to the active safeguard configuration

This interaction record is not the interaction itself. It is a concise cryptographic attestation of the interaction's governance properties. The content stays in the application layer; the governance evidence goes into the audit trail.

Zero-Knowledge Proof of Authorisation

A ZK proof binds the interaction record to the authorisation registry without requiring the auditor to examine either the interaction content or the full registry. The circuit proves: "There exists an authorisation record in the registry such that the tool used in this interaction matches the authorisation record, the data classification was within the authorisation scope, and the safeguards were active." The auditor verifies the proof; they do not access the interaction content or the employee's identity. The proof is approximately 192 bytes (Groth16 SNARK) and verifies in milliseconds.

For an AI governance audit, this means a regulator can be shown proof that every AI interaction in a given period was conducted with an authorised tool under active controls, without the organisation disclosing the content of its employees' AI-assisted work. This satisfies the EU AI Act's Article 12 accountability requirements, the GDPR's data minimisation principle (Article 5(1)(c)), and the PLD's evidentiary requirements, simultaneously.

Post-Quantum Binding with ML-DSA-65

The ZK proof transcript is signed with ML-DSA-65 (NIST FIPS 204) to provide post-quantum binding. A quantum adversary who breaks the classical cryptography underlying the ZK proof system cannot retroactively forge a signed proof transcript without also breaking ML-DSA-65 under the MLWE assumption. For audit records that must remain valid for 10 years, this matters: a compliance record signed in 2026 with ECDSA may be retroactively forgeable by a quantum adversary before its regulatory retention period expires. ML-DSA-65 signed records remain tamper-evident for the full retention period under current post-quantum security estimates.

SECTION 08

Merkle-Anchored AI Interaction Records

A large organisation may generate millions of AI interactions per day. The audit architecture must handle this at scale without creating an unmanageable volume of cryptographic signatures. The solution is a Merkle-anchored batch attestation, described in full in WP-032: Sublinear Post-Quantum Attestation.

In brief: interaction records are batched (by time period or count), assembled into a Merkle hash tree, and the root of the tree is signed once with ML-DSA-65. Individual interaction records are verified via an inclusion proof of approximately O(log n) hash values. The storage efficiency is dramatic:

SCALE	PER-RECORD ML-DSA-65 SIGNATURES	MERKLE-ANCHORED (ONE ROOT)	REDUCTION
100K interactions/day	314 MB/day signatures alone	3,293 bytes (root)	>99.9%
1M interactions/day	3.14 GB/day	3,293 bytes (root)	>99.9%
10 years at 1M/day	11.4 TB signatures alone	~12 MB roots + records	>99.9%
Verify one interaction (1M batch)	1 ML-DSA verify	20 SHA-256 hashes + 1 ML-DSA verify	Equivalent

The Merkle-anchored record satisfies all five requirements from Section 6: it is immutable (hash-tree structure detects any alteration), complete (every interaction is a leaf), attributed (user identity is hashed into each leaf), content-binding without content disclosure (the leaf is H(interaction content) not the content), and compatible with regulatory retention (roots are published to an external immutable registry at anchoring time, preventing retroactive substitution).

The signed root can be published to a public blockchain, a qualified timestamping service (RFC 3161), or a regulatory registry, providing an external, independently verifiable record that the signed batch existed at a specific time. This matters for PLD defence: if an incident occurs, the organisation can demonstrate that their AI governance records predated the incident and were not constructed retrospectively.

SECTION 09

Shadow AI at Machine Speed: The Agentic Dimension

The shadow AI problem is serious when it involves human employees making individual AI-assisted decisions. It is a different category of problem when it involves autonomous AI agents making thousands of decisions per minute

without human review.

Shadow agentic AI is already the dominant emerging risk. According to the Gravitee State of AI Agent Security 2026 report:

- 80.9% of technical teams have AI agents in active testing or production.
- Only 14.4% of those agents went live with full security and IT approval.
- More than half of all enterprise AI agents operate without consistent security oversight or logging.

Clifford Chance's February 2026 analysis of agentic AI liability identified the core problem: agents take actions. They do not just generate suggestions for humans to review. An agent that incorrectly authorises a supplier payment, misprices a product, sends a misleading customer communication, or makes a biased employment decision has caused harm before any human becomes aware it happened. The AI vendor's contract will disclaim liability for the harm. The organisation's policy about AI use will not prevent it. Only a real-time audit trail that captures the agent's decision at the moment it was made, and proves that the agent was operating within authorised parameters at that moment, provides the evidentiary basis for defence or damages recovery.

The Agent Authorisation Chain

For agentic AI, the audit trail requirement extends beyond tool authorisation to action authorisation. It is not sufficient to prove that the agent was an authorised tool. It must be demonstrable that each specific action the agent took was within the scope of what the agent was authorised to do. This requires a proof chain: each agent action record references the authorisation scope, and the authorisation scope is signed by the authorising officer. The chain is: human officer authorises agent scope; agent action references scope; action record is cryptographically bound to scope record; scope record is cryptographically signed. Any action outside the scope is detectable because it cannot produce a valid proof chain.

This architecture is described in detail in WP-015: Cryptographic Audit Trails for Autonomous AI Agents. The key addition in the shadow AI context is that the agent itself may be a shadow agent: deployed without IT approval, connecting to company systems via personal API keys or OAuth tokens,

operating without any sanctioned logging infrastructure. In this case, even the detection of the agent's existence may depend on data access anomaly monitoring rather than any compliance control.

SECTION 10

Regulatory Mapping

Shadow AI governance interacts with multiple regulatory frameworks simultaneously. The following table maps the specific requirements that a shadow AI audit trail must satisfy.

REGULATION	KEY SHADOW AI REQUIREMENT	WHAT THE AUDIT TRAIL MUST PROVE	RETENTION
EU AI Act (2024/1689) Article 12	High-risk AI systems must maintain logs sufficient to reconstruct decisions	Model version, data inputs, active safeguards at decision time	10 years
EU Product Liability Directive (Dec 2026)	AI tools are products; non-compliance with EU AI Act = defective product	Tool was authorised, risk-assessed, compliant at point of use	Applicable limitation period
GDPR Article 5 (data minimisation)	AI processing of personal data must be limited to what is necessary	Data classification was checked; access was within authorised scope	Until subject to erasure request
GDPR Article 30 (records of processing)	Records of processing activities must be maintained	AI tool is identified as a processor; processing activity is logged	Ongoing
DORA (EU 2022/2554) Article 12	ICT systems must be fully documented and auditable	AI tool classified as ICT system; security events logged immutably	5 years
HIPAA 45 CFR 164.312	Access controls and audit trails for PHI systems	PHI accessed via AI tool was within authorised access scope	6 years
Colorado SB 24-205 (effective 2026)	High-risk AI in consequential decisions must be auditable for bias	Model version, training data provenance, bias test results at deployment	3 years minimum
EU AI Liability Directive (in progress)	AI-caused harm creates rebuttable presumption of causation	AI tool in scope of harm was operating within design parameters	Applicable limitation period

The common thread across all these frameworks is the same: they require evidence of what the AI system actually did, not what the policy said it should do. A policy satisfies none of these requirements. An immutable, cryptographically authenticated audit trail satisfies all of them.

SECTION 11

From Policy to Proof: Implementation

Moving from a policy-based shadow AI governance posture to a proof-based one does not require replacing existing infrastructure. It is an overlay architecture: the proof layer sits between existing AI tools and existing data systems, capturing governance evidence at the point of interaction. The following four-stage process is how AffixIO approaches this migration for organisations.

Stage 1: Inventory (Weeks 1-4)

Identify every AI tool in use across the organisation: sanctioned, unsanctioned, and embedded. Sources include: network proxy logs, browser extension inventories, OAuth token registries (for agent integrations), vendor AI-feature disclosures, and employee self-declaration surveys. The inventory will likely reveal a significantly larger AI surface than the security team expected. 60-70% of organisations discover AI tools in use that they were unaware of during this stage. The output is a prioritised list of tools by data exposure risk: tools processing personal data, financial data, legal matter information, and intellectual property are the highest priority.

Stage 2: Classify and Authorise (Weeks 3-8)

For each AI tool identified, make an explicit authorisation decision: authorised as-is; authorised with controls (data classification limits, safeguard requirements); under review; or blocked. Create signed authorisation records for each approved tool. The signed records become the reference data for the audit trail: when an interaction is recorded, it references the authorisation record active at that time. Tools authorised before December 9, 2026 should be re-reviewed for PLD compliance before that date.

Stage 3: Instrument (Weeks 6-16)

Deploy instrumentation at the integration points where AI tools are used. For sanctioned enterprise tools (Microsoft 365 Copilot, GitHub Copilot, etc.), instrumentation hooks into the tool's audit log APIs. For browser-based consumer AI tools, network-level instrumentation captures interaction metadata (not content). For agentic AI integrations, instrumentation wraps each API call with a signed action record. The instrumentation outputs interaction records (not transcripts) to the Merkle-anchored audit trail.

Stage 4: Attest and Publish (Ongoing from Week 10)

Batch interaction records are assembled into Merkle trees at configurable intervals (typically every 60 seconds for real-time applications, up to daily for archival). Each batch root is signed with ML-DSA-65 and published to an external anchoring service. The publication timestamp establishes that the compliance records existed before any incident, providing the retrospective evidentiary foundation that the PLD requires.

> **Time to first proof:** In AffixIO's deployment experience, the time from initial engagement to first cryptographic proof of AI tool authorisation is typically 2-4 weeks for the highest-risk tool categories, with full coverage of the AI tool inventory achievable within 90 days. This is significantly faster than the PKI migration timescales (5-10 years) that apply to post-quantum certificate infrastructure, because the ZK proof overlay does not require replacing existing authentication systems.

SECTION 12

Conclusion

Shadow AI is not a problem you can solve with a better policy. The data is consistent: most organisations have policies; most organisations have not audited compliance with those policies; most breaches occur in organisations that lacked the access controls that their policies said they should have. The structure of the problem guarantees policy failure, because the tools

proliferate faster than policies can be written, detection is retrospective, and detection-based approaches create records in mutable systems that will not survive regulatory scrutiny.

The December 2026 deadline imposed by the EU Product Liability Directive is a forcing function. After December 9, every AI tool placed into service is a product subject to no-fault liability if it causes harm. The only defence is documented evidence that the tool was authorised, compliant, and operating within sanctioned parameters at the moment of harm. Policies do not produce that evidence. Audit trails do, provided they are immutable, attributed, content-bound without content disclosure, and retained for the applicable regulatory period.

The architectural shift required is from detecting unauthorised AI use after the fact to generating cryptographic proof of authorised AI use at the moment it occurs. This requires four elements that a policy never provides: runtime instrumentation, immutable storage, cryptographic binding, and external anchoring. The ZK proof layer adds the privacy property: auditors verify authorisation without accessing the content of AI interactions. The ML-DSA-65 signing layer adds the future-proofing property: compliance records remain tamper-evident through the full regulatory retention period, including the window in which quantum computers may become cryptographically relevant.

The cost of not building this infrastructure is, in each of the following terms, now quantifiable:

- **Financial:** Shadow AI adds approximately $670,000 to the average breach cost and 10 additional days of containment time (IBM, 2025).
- **Legal:** PLD strict liability from December 2026; CISO personal liability risk at 78% concern rate.
- **Regulatory:** EU AI Act Article 12 penalties; GDPR Article 83 fines of up to 4% of global annual turnover for processing violations.
- **Reputational:** The Samsung incident cost months of remediation and significant restrictions on legitimate AI use. The next high-profile shadow AI breach will cost more.

ISACA said it plainly in 2026: "Policy is intent. Proof is governance." The organisations that take this seriously before December 2026 will be in a demonstrably different legal position than those that do not. AffixIO provides one component of the infrastructure needed to make that shift — the cryptographic proof layer — as part of a broader ecosystem of AI governance tools. The policy layer remains the organisation's responsibility; the proof that the policy was followed is what we make verifiable.

> Related AffixIO whitepapers: WP-015: Cryptographic Audit Trails for Autonomous AI Agents covers the agentic AI proof chain in depth. WP-032: Sublinear Post-Quantum Attestation covers the Merkle+ML-DSA-65 storage architecture. WP-002: Post-Quantum Attestation covers ML-DSA-65 production deployment. WP-003: The Proof-Not-Log Paradigm covers the foundational case for cryptographic evidence over mutable logs.

FREQUENTLY ASKED

Shadow AI Governance: Common Questions

What is shadow AI and why is it a governance problem?

Shadow AI refers to AI tools used within an organisation without IT, security, or compliance approval. It is a governance problem because: 65% of shadow AI incidents expose personal data; 97% of AI-related breaches occur in organisations without access controls; and from December 2026, the EU Product Liability Directive classifies AI systems as products subject to no-fault liability, meaning organisations that cannot document authorised use face direct legal exposure.

How widespread is shadow AI in 2026?

Extremely widespread. 67% of employees use AI tools at work; only 18% of companies have AI security policies. 29% of employees use unsanctioned AI agents for work tasks. Shadow AI-related data loss increased nearly four times year-on-year. Only 13% of organisations combine written policies with regular audits.

What is the difference between a shadow AI policy and an audit trail?

A policy states what should happen. An audit trail proves what actually happened. ISACA (2026) states this directly: "Policy is intent. Proof is governance." A policy cannot be produced as regulatory or legal evidence the way an audit trail can. The EU AI Act, DORA, HIPAA, and the EU Product Liability Directive all require evidence of actual system behaviour, not statements of intended behaviour.

What does the EU Product Liability Directive mean for shadow AI after December 2026?

The PLD, effective December 9, 2026, classifies AI systems as products under strict (no-fault) liability. If an AI tool causes harm, the deploying organisation is liable even without negligence. The PLD links defectiveness to non-compliance with the EU AI Act. An unsanctioned AI tool (shadow AI) is non-compliant by definition; non-compliance means defectiveness; defectiveness means liability. The only defence is documented evidence of authorised, compliant use at the point of the relevant interaction.

WP-015: Agentic AI Governance | WP-003: Proof Not Log | WP-032: Sublinear Attestation

▶ About
▶ Solutions
▶ Legal
▶ Trust & Security

Contact

truth layer | yes | no | proof