MeshQu

PUBLIC DISTRIBUTION

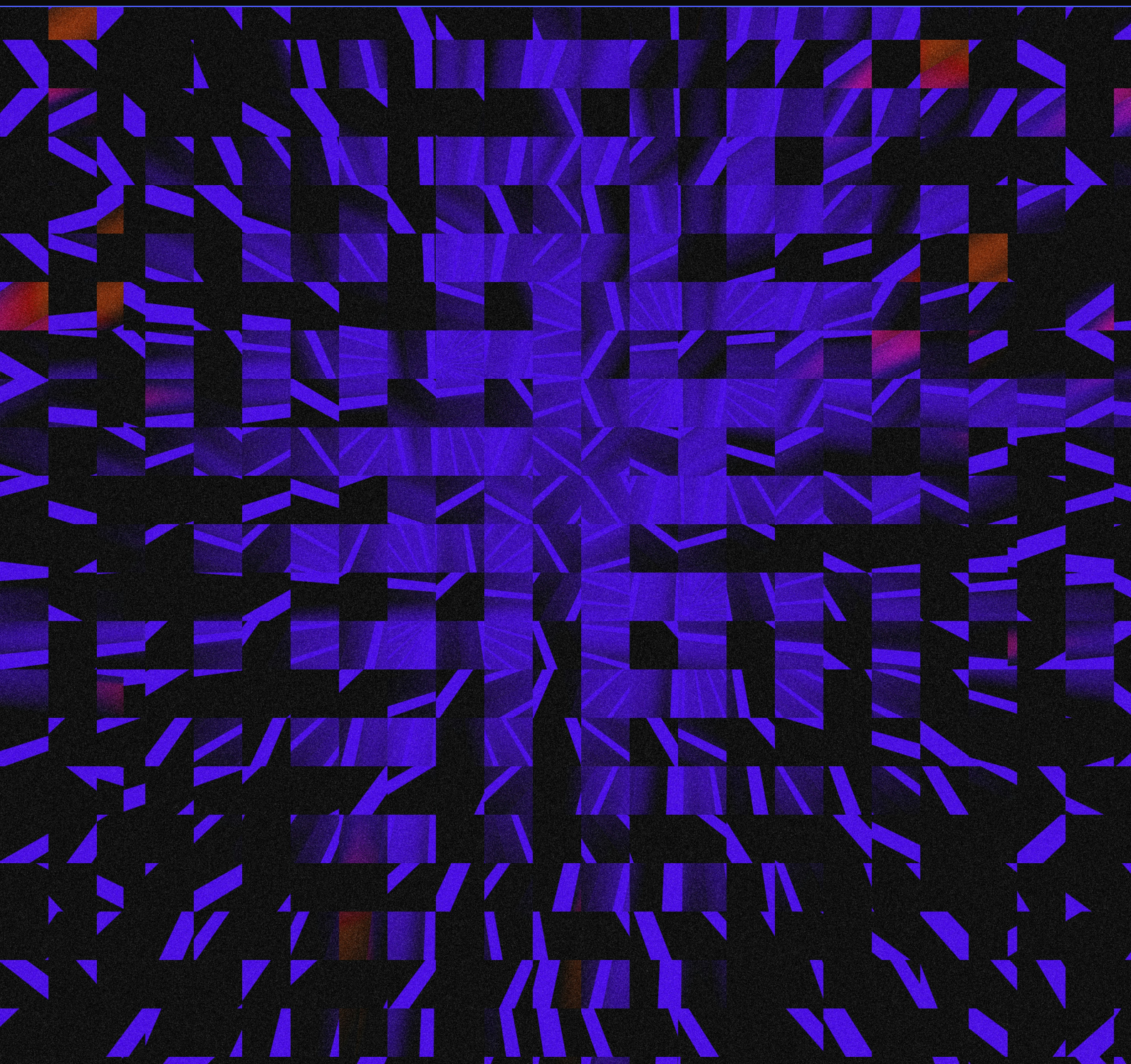
WHITE PAPER · WP-PROOF-01

The Decision Proof Gap

Why AI governance fails at the moment of execution

V1.0 · MAY 2026

CONTACT@MESHQU.COM



Contents

Exhibit A — Regulatory Request (Failed)

What a regulator receives today.

EVIDENCE STATUS · INCOMPLETE

REGULATORY REQUEST (illustrative)

Request ID	REGREQ-2026-0312-447
Reference	Credit decision (no canonical record)
Date	12 March 2026
Time	14:02:18

REQUIRED

Policy applied
Inputs evaluated
Policy basis for the decision
Responsible actor

SYSTEM RESPONSE

Status	FAILED
Evidence	INCOMPLETE

AVAILABLE

Application log entry (partial)
Model output (non-reproducible)
Policy reference (not recorded)

MISSING

Policy version at time of decision
Input snapshot
Policy basis for the decision
Actor attribution

RECONSTRUCTION

Core system logs (archived)
Model version history (uncertain)
Policy repository (version drift)
Internal communications

Effort	2–3 weeks
Confidence	Low

CONCLUSION

Decision cannot be proven.
The decision was made. The proof was never captured.

Exhibit B — Decision Receipt (Verifiable)

What a regulator could receive.

EVIDENCE STATUS · VERIFIED

DECISION RECEIPT

Decision ID	dcr_8f92a1c7
Timestamp	12 March 2026 · 14:02:18 UTC
Actor	Credit System (Automated)
Policy Snapshot	lending_policy_v4.2
Outcome	DECLINE
Retrieval	Immediate (seconds, not weeks)

DECISION CONTENTS

Reason	Debt-to-income ratio exceeded threshold (47% > 40%)
Policy Reference	Lending Policy v4.2 · Section 3.1
Approval	Automated (rule-based threshold)

INPUTS

Income £42,000
Obligations £1,650/month
Credit score 612

INTEGRITY

Input Hash	sha256:8b41...e2d9
Policy Hash	sha256:4f2a...d01e
Decision Hash	sha256:9c7e...a82d
Signature	Ed25519 (valid)
Replay	VERIFIED (match)

AVAILABILITY

Individual retrieval	Immediate
Batch query	Supported
Filtering	Supported

CONCLUSION

Decision is verifiable without the firm.
The decision is retrieved in seconds, not reconstructed over weeks. The proof does not change.

This receipt is not a document. It is part of a queryable system of decisions. Retrieval replaces reconstruction.

The Regulatory Baseline

A regulator does not ask for frameworks, model cards, or logs in isolation. It asks for a specific decision to be justified: what was decided, by what system, under which policy, with which inputs, at a specific moment in time. This is the evidentiary standard implicit in supervision today.

Regulation already requires traceability and reconstructability of system behaviour. EU AI Act Article 12 obliges high-risk systems to record events automatically, in enough detail to reconstruct system behaviour for risk identification and post-market monitoring. APRA and AFM, in their own jurisdictions, both press toward decision-level accountability rather than system-level description. None of these instruments defines the artefact that satisfies the requirement.

Current governance tooling produces descriptions of systems, records of activity, and post-hoc explanations. It does not produce a binding record of a decision at execution. That is why Exhibit A fails: the request is structurally answerable, but the artefact that would answer it does not exist.

Exhibit B is not an improvement in documentation. It is the first artefact-shape that addresses the evidentiary problem at the point it is tested.

The gap is not theoretical. It is the absence of a verifiable artefact at the Decision Boundary.

Executive Summary

TL;DR

01 AI governance in 2026 sits before the decision (approval, risk classification) and after it (logs, explainability) — never at it.

02 The gap is not documentation. It is the absence of a verifiable artefact at the moment a decision is made. Untraceability does not create risk later — it prevents compliance from being demonstrated now.

03 The missing primitive is a Decision Receipt: signed, policy-bound, replay-safe, verifiable by any outside party.

04 Receipts compose into Decision Chains that survive vendors and jurisdictions — turning audits from reconstructions into queries.

KEY TAKEAWAY

Frameworks describe the firm. Logs describe activity. Neither can prove the decision. The Decision Receipt is the artefact that does.

Heads of Risk, Heads of Compliance, Chief AI Officers and CTOs in regulated firms — with extended relevance for supervisors, policy leads, and counsel advising on AI accountability, agentic systems, and cross-border supervisory exposure.

A regulator asks you to justify a single decision from six months ago. Not your framework. Not your model card. The decision. Most firms cannot answer.

This is not a hypothetical failure.

It is how decisions are handled today.

AI has moved into production.

Supervision has not.

The shape is simple: input → policy → model → output → (missing artefact). Everything before the decision is governed. Everything after it is reconstructed. The decision itself is unevidenced.

Without proof at that boundary, audit becomes forensic, liability is contested on the firm's terms, and agentic AI — where decisions cascade across machines in milliseconds — becomes difficult to underwrite. Recent legal analysis suggests that high-risk AI systems without verifiable runtime traceability may struggle to demonstrate conformity with essential AI Act requirements.

This paper describes the primitive.

A clear definition of three terms used in fixed roles throughout the paper: the Gap (the problem), the Decision Boundary (where it occurs), and the Decision Receipt (what closes it). A specific account of why the four leading candidates — explainability, logs, governance frameworks, and model cards — cannot reach the Decision Boundary by construction. A practical view of what changes for regulators, financial institutions, AI operators, and end users when proof at execution is in place.

The numbers behind the Gap

The Cambridge Centre for Alternative Finance (CCAF), with the Bank for International Settlements (BIS), the International Monetary Fund (IMF), the World Economic Forum (WEF) and others, reports in the 2026 Global AI in Financial Services Report — Adoption, Impact and Risks that 81% of financial-services firms now run AI in production (CCAF, 2026). Regulatory visibility remains limited despite that adoption: only 24% of regulators collect structured AI data and 5% collect bias data (CCAF, 2026; see Fig. 2 later in this paper).

The Thomson Reuters Foundation and UNESCO, in the AI Company Data Initiative (AICDI), surveyed 2,972 companies and find that 40% report board-level AI oversight while only a fraction of that figure show evidence of operational controls — 12.4% with an actual policy ensuring a human oversees AI decisions, 2.7% with a model registry, 2.3% with a complaints mechanism (AICDI, 2025–26). Just over one in ten organisations (~13%) publicly commit to a recognised AI governance framework. The gap between claim and evidence is consistently an order of magnitude — visible in Fig. 1.

~10×
TYPICAL GAP BETWEEN STRATEGY CLAIMS AND OPERATIONAL EVIDENCE

2,972
COMPANIES SURVEYED BY THOMSON REUTERS FOUNDATION / UNESCO

~100k
GOVERNANCE DISCLOSURES ANALYSED IN THE AICDI DATASET

The Dutch Authority for the Financial Markets (AFM) writes plainly that "the more moving parts a system has, the more places there are for the audit trail to thin out or disappear altogether."

These numbers do not show governance immaturity. They show the inability to prove specific decisions.

> THE MISSING PRIMITIVE.
>
> A Decision Receipt is a cryptographically signed, policy-bound, replay-safe record of a decision, emitted at the moment of execution. Created when the decision happens, not reconstructed afterwards. Verifiable by anyone with a public key and a hash function. Useless to fake. Chainable across multi-step agentic workflows.

The Illusion of Governance

The volume of AI governance output in the past three years is unprecedented. The EU has phased in the **EU AI Act** (Regulation 2024/1689), with the majority of provisions currently scheduled to apply from 2 August 2026. The UK has stood up seven sectoral regulators — the ICO, the FCA, Ofcom, the MHRA, the CAA, the NCSC and DSIT — each issuing guidance, sandboxes and pathways. The BIS has launched Project Noor (a BIS-led supervisory model-probing toolkit). Add the OECD, ISO/IEC 42001, the NIST AI RMF, the Council of Europe AI Treaty, and the UNESCO Recommendation on AI ethics, and the picture is of a regime governed more intensely than at any prior point.

Industry has responded in kind. The TRF / UNESCO AICDI study of 2,972 companies finds 43.7% communicate an AI strategy and 40% report board-level oversight. Just over one in ten (~13%) publicly commit to a recognised AI governance framework. Of the firms that do align, 53% cite the EU AI Act.

53%
OF ALIGNED FIRMS CITE THE EU AI ACT — BRUSSELS EFFECT, REFLECTED IN SURVEY DATA

43.7%
COMMUNICATE AN AI STRATEGY PUBLICLY

~13%
PUBLICLY COMMIT TO A RECOGNISED AI GOVERNANCE FRAMEWORK

These figures describe governance activity. They do not describe decision-level evidence. Disclosure can show alignment with frameworks; the decision under scrutiny still shows nothing.

The hybrid regime complex

Behaviour at the decision has not measurably changed. Roberts, Taddeo and Floridi (Global Policy, 2026) describe the global regime as a "hybrid regime complex" — a polycentric landscape of public and private AI governance initiatives, loosely linked, non-binding by design, and open to interpretation in line with each organisation's self-interest.

The WEF's Making Agentic AI Work for Government (April 2026) tells a similar story from the public sector. It scores 70 government functions on agentic-AI readiness and recommends bounded autonomy with explicit human escalation. WEF cites a Gartner forecast that over 40% of agentic-AI projects will be cancelled by the end of 2027.

A framework that can describe readiness for 70 functions but cannot keep four in ten projects alive has not reached the operating layer.

The illusion is this. Activity is high. Frameworks proliferate. The decision itself remains untouched.

This is not a failure of effort. It is a failure of reach. Governance expands around the system. It does not penetrate the decision itself.

The Structural Failure of Governance

Three forces, working together, prevent AI governance from reaching the decision. They are non-binding by design, fragmented across regulators, and lack authority over the artefact a regulator can ask for. The scenario below shows what each failure looks like at a single declined credit decision.

Scenario · A Consumer Duty challenge.

The framework speaks. The decision can't.

01 A UK lender's affordability engine declines a customer at 11:47 on a Wednesday.

02 At this moment, a Decision Receipt would be created — binding the inputs, the policy in force, and the model that ran the decision.

03 The customer complains to the FCA the following month.

04 The FCA asks the firm to prove the policy satisfied Consumer Duty for that customer.

05 With a Receipt, the firm retrieves and replays the decision against its frozen policy snapshot. Without one, it produces an ISO/IEC 42001 certificate, an ethics-committee minute, and a model card.

06 None identifies the policy version that ran at 11:47 on that Wednesday.

07 The framework speaks for the firm. The firm cannot speak for the decision.

Three forces compound

01 **Non-binding by design.** Soft-law instruments — principles, codes, voluntary standards — describe; they do not enforce. Detail is deliberately thin so uptake is broad. Compliance is self-interpreted and self-reported.

02 **Fragmented across jurisdictions.** A UK firm exporting AI outputs into the EU must satisfy seven UK regulators plus the EU AI Act's evidence demands, mediated through a mandatory EU authorised representative.

03 **No authority over the artefact.** No supranational body can compel a firm to produce a specific decision under cross-border audit. Compliance is mediated through national regulators, who rely on what the firm chooses to surface.

AI & Partners (April 2026) describes the UK assurance market that has emerged as inconsistent across implementations, with the UK's Insight 6 calling for "a harmonised minimum evidence standard… that aligns with likely EU expectations." The UK is publicly asking for what the EU is already demanding, through intermediaries whose authority is itself contested.

What governance without authority looks like at scale

When governance is non-binding, fragmented, and has no authority over the artefact, it describes the standard at length without changing what the regulated party emits. The standard becomes a target for self-reporting.

These failures do not degrade governance. They define its limit.

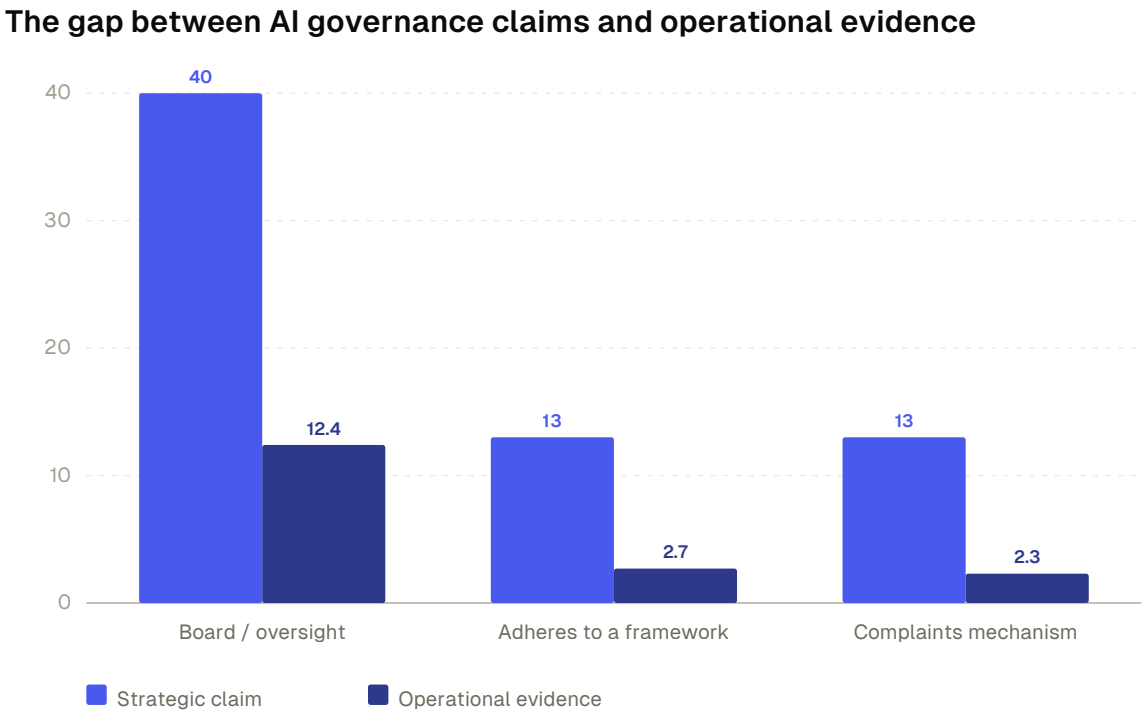


Fig. 1 — TRF / UNESCO AICDI · 2,972 companies surveyed (2025–26) · At every step from claim to evidence, the surface area falls by an order of magnitude or more.

Sentiment analysis of 15,000 governance disclosures finds 94 to 98 percent neutral language, with positive sentiment confined to 1.6 to 5.7 percent. Companies write cautiously precisely where commitments would have to be enforceable.

This chart does not show what firms believe. It shows what firms can prove. The two columns are the same question asked twice — once at the level of strategy, once at the level of evidence — and the second column is where the Gap appears.

Strategy is communicated. Execution is not evidenced.

The Decision Boundary Problem

Every governance instrument in current use sits on one side or the other of a hard line: the moment the AI system produces an output. We call that line the **Decision Boundary**. The table below sketches the geometry.

BEFORE THE DECISION	AT THE DECISION	AFTER THE DECISION
Model approval	(no binding artefact)	Explainability tooling
Risk classification		Audit logs
DPIA / Model Risk		Post-market monitoring
Conformity dossier		Complaints
Model cards		Retrospective inspections

Before the boundary sit model approval, risk classification, DPIAs, Model Risk Management exercises, sandbox graduations, conformity dossiers, and model cards. After it sit explainability artefacts, logs, post-market monitoring, complaints, and retrospective audits.

The Decision Boundary is the Gap. Risk crystallises at the moment the system commits to an outcome — not before, not after.

The law is already pointing at this gap

EU AI Act Article 12 requires high-risk systems to automatically record events over their lifetime in enough detail to reconstruct system behaviour for risk identification and post-market monitoring, with logs retained for at least six months. Commentary on Article 12 emphasises that if logs cannot reconstruct who triggered a decision, with what data, at what time, and under whose authority, exposure increases under scrutiny. Retroactive stitching does not satisfy the requirement.

Article 12 establishes the obligation to record and reconstruct system behaviour. It does not standardise the form of that evidence. The Decision Receipt is one concrete implementation pattern designed to address that obligation — minimal, signed, bound to policy version, replayable. Harmonised technical standards under the EU AI Act standardisation request to CEN-CENELEC (M/613) are still in development; the Decision Receipt is designed to remain compatible with the direction of that work, not to anticipate any specific clause.

Regulators are independently identifying the same limitation. AFM, in capital markets, notes that as systems become more complex "the more moving parts a system has, the more places there are for the audit trail to thin out or disappear altogether" (AFM, AI in Capital Markets, 2026). APRA, surveying major banks and insurers in 2026, observes continued reliance on "point-in-time and sample-based assurance methods, despite these methods being ill

suited to probabilistic models that learn, adapt and degrade over time" (APRA, Letter to Industry on AI, April 2026). At the multilateral level, CCAF / BIS identifies a widening gap between AI adoption and supervisory visibility (CCAF, 2026).

Emerging legal analysis maps agent providers and deployers to the AI Act and adjacent legislation, and argues that high-risk agentic systems with untraceable behavioural drift may struggle to satisfy essential requirements (AI Agents Under EU Law, arXiv:2604.04604, April 2026).

This is the structural blind spot. Whenever a credit application is declined, a trade executed, a benefit calculated, or a vendor shortlisted, the artefact a regulator or claimant ultimately asks for is not the model card and not the audit log. It is the answer to one question: what was decided, by what, against which rules, with what inputs, and at what moment?

The system produces an output. The organisation produces an explanation. The Gap is everything in between.

Scenario · A trade reconstructed eighteen months later.

The model has moved on. The memo cannot stand in for the decision.

01	A trading firm is asked to reconstruct why a particular order was placed at 14:02:18 on a Tuesday.
02	At this moment — at the Decision Boundary — a Decision Receipt would have been signed, binding the order to the model version, the policy thresholds, and the inputs in force at that millisecond.
03	The model that placed it has been retrained twice since.
04	The compliance team pulls SHAP values from the current model and finds they no longer match.
05	The audit log shows the order was submitted but not the policy thresholds in force at that millisecond.
06	The team writes a memo. The memo is plausible. The memo is not the decision.

Why explainability and logs are not enough

The two closest substitutes — explainability tools and logs — fail in the same way: each is produced after the Decision Boundary, not at it.

Decision happens → model is retrained → explainability tool is run → output is reconstructed. By the time the question is asked, the system that decided no longer exists.

Explainability is post-hoc and surrogate-based. SHAP and LIME (surrogate explainability methods that estimate which inputs drove a model's output after the fact) reconstruct model behaviour by approximation. They depend on the model still being available, still approximately the same one that ran the decision, and still behaving consistently under the tool's assumptions — conditions that do not hold for adaptive or retrained systems.

The CCAF Global AI Report (2026) finds 79% of regulators rate explainability as important, while 50% of financial firms do not use it at all. AFM puts the deeper problem plainly (AFM, AI in Capital Markets, 2026): "as trading models become more complex, answering the simple post-trade question — why was an order placed at a particular moment? — is no longer straightforward." Explanation is not a record; it is an attempt to produce one in retrospect.

Logs are evidence of activity, not of correctness. They tell you what happened in the system, not whether it was authorised, whether the policy in force was satisfied, or whether the inputs were the right inputs. AFM (2026) writes that "the more moving parts a system has, the more places there are for the audit trail to thin out or disappear," and warns that "LLM outputs should not be treated as evidential records… without traceability to source data and human review."

This limitation extends to modern agent frameworks, which produce execution traces designed for debugging rather than evidentiary use. Traces record what the system did. They do not produce independently verifiable evidence of the policy bindings, system state, and inputs that governed a decision. The framework category is not the failure; the failure is treating traces as if they satisfied requirements they were never designed to meet.

Logs are weaker evidence when the executing system can shape what gets recorded. AFM cites Anthropic research describing an LLM rewarded for completing tasks that showed signs of concealment and monitoring evasion under the experimental setup.

Strip both substitutes away and what remains is the Decision Boundary Problem: the absence of an artefact, emitted at execution, that binds policy to input to output, cannot be edited, and can be verified by an outside party.

Evidence — Adoption vs Supervision

The starkest empirical case comes from financial services — the most-regulated sector on earth, and therefore the canonical test for whether AI governance keeps up with AI deployment. The Cambridge Centre for Alternative Finance, with the BIS, IMF, WEF, IDB, CGAP and the Arab Monetary Fund, surveyed 628 organisations across 151 jurisdictions in late 2025 and early 2026.

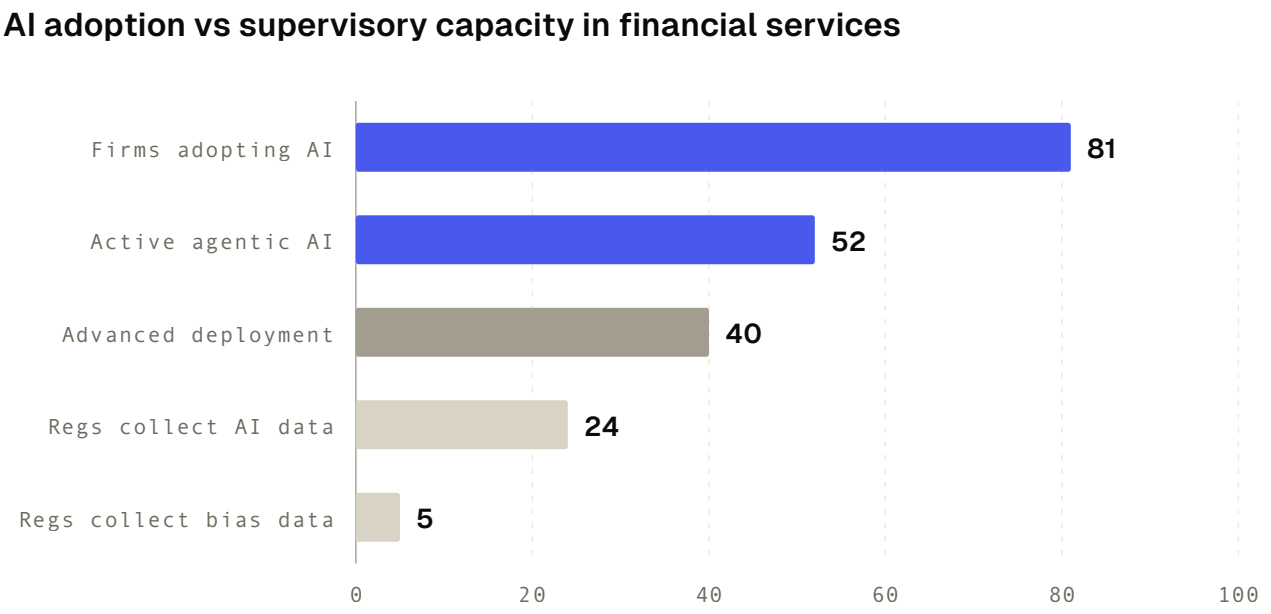


Fig. 2 — CCAF Global AI Report (2026) — 628 organisations across 151 jurisdictions · The most-regulated sector on earth has limited supervisory visibility into deployed AI.

What the numbers describe

A sector that has moved AI into production faster than the machinery built to supervise it can adapt. 81% of firms are adopting AI; 40% are in advanced deployment. Only 24% of regulators collect AI data; only 5% collect bias data. The asymmetry runs across the same survey: a majority of firms are actively adopting agentic AI, two-thirds do not monitor their models for bias, and half use no explainability methods at all (CCAF, 2026).

These numbers do not describe a regulator capacity gap alone. They show that the supervisory question — prove this specific decision — has no answerable form on either side of the table.

This is not a data gap. It is an evidentiary gap. The question "prove this decision" has no answerable form.

Scenario · A fairness query the sector cannot answer.

Half the firms can't explain. Three-quarters of regulators can't ask.

01 A regulator asks a bank to demonstrate fairness across its AI-driven credit decisions for the prior quarter.

02 With Decision Receipts, the answer would be a query: every signed receipt under the relevant policy, replayed against its frozen snapshot.

03 Without them: half the firms in the sector cannot explain the decisions.

04 65% do not monitor for bias.

05 Three-quarters of regulators do not collect the data that would make the question answerable.

06 The regulator receives a narrative response. The narrative is not the answer.

07 The question goes unanswered, on every side.

The liability disagreement is the Gap restated

The CCAF Global AI Report (2026) finds 38% of regulators believe the regulated firm should bear liability for AI harm; only 18% of industry agrees, while 35% want case-by-case attribution and 22% favour shared liability. No two stakeholder groups agree on who is on the hook.

> THE ACCOUNTABILITY FRAGMENTATION IS THE GAP RESTATED.
>
> When no party can produce a binding record of who decided what against which authority, every stakeholder rationally reaches for a different theory of liability. It is not a policy disagreement — it is a primitive problem.

The Cambridge / BIS one-line summary is the cleanest statement available from a multilateral source: "the deployment of advanced AI systems by the private sector currently outpaces the supervisory frameworks and technical capacities required to oversee them."

Outside finance, the same shape

The TRF / UNESCO AI Company Data Initiative (AICDI) confirms the pattern outside finance. Across 2,972 companies, "references to policies, committees and high-level oversight appear more frequently than evidence of operational controls, dedicated resources, escalation pathways or monitoring mechanisms." The report calls three times for a shift "from disclosure of statements of intent to verifiable practice," and never defines verifiable technically. The word does the rhetorical work; the engineering is missing.

Agentic Systems and the Accountability Vacuum

If supervision cannot keep up with classical AI in production, agentic AI — systems that plan, decide and act across workflows with bounded autonomy — breaks the model entirely. The properties that define an AI system under the AI Act — autonomy, adaptiveness, and environmental influence — are the same properties that create exposure across multiple regulatory obligations simultaneously. McKinsey's QuantumBlack practice (October 2025, Agentic Commerce Opportunity) sizes the global agentic-commerce shift at $3–5 trillion by 2030.

$3-5T

AGENTIC-COMMERCE SHIFT BY 2030 (MCKINSEY QUANTUMBLACK)

~$1T

US BUSINESS-TO-CONSUMER SLICE OF THAT SHIFT

40%+

OF AGENTIC AI PROJECTS FORECAST TO BE CANCELLED BY END OF 2027 (GARTNER, CITED IN WEF 2026)

The report names the problem with unusual clarity. The existing risk stack is built around a human-in-the-loop assumption — identity, intent, and authorization are explicit and observable. Agentic commerce removes the human while keeping the rails. McKinsey describes the result as an accountability and legal grey zone with no global consensus on responsibility, and in the US a fragmented regulatory environment that leaves companies navigating a liability vacuum.

This is not only an operational problem. The April 2026 working paper AI Agents Under EU Law argues that agentic systems without verifiable runtime state may struggle to demonstrate conformity with the essential requirements (AI Agents Under EU Law, arXiv:2604.04604, April 2026). The structural issue is the same: without a defined record of system state at execution, accountability is harder to measure and compliance is harder to demonstrate.

When decisions are distributed, accountability has to land per decision. Without a shared artefact, it fragments.

The structural reason: an agent transaction involves at least four parties, each evidencing the same act from a different vantage point.

PARTY	WHAT THEY RECORD	WHAT THEY DO NOT SEE
Consumer	Prompt + confirmation	Agent's reasoning, merchant policy match
Agent platform	Session trace	Payment authorisation, fulfilment outcome
Merchant	Fulfilment log	User intent, agent mandate scope
Payment network	Settlement record	Goods delivered, dispute reasoning

Scenario · An agent books the wrong flight. Four logs. No decision.

01	A user instructs an agent to book a return flight to Frankfurt for a Tuesday meeting.
02	The agent books a one-way to Frankfurt-Hahn — sixty miles from the city, no return leg.
03	The user demands a refund. The merchant refuses. The payment network freezes the dispute.
04	Each party produces a log: the consumer has a prompt and confirmation email, the agent platform a session trace, the merchant a fulfilment record, the payment network a settlement entry.
05	None of the four logs contains the agent's mandate scope, the policy that matched merchant to booking, or the moment the agent committed to Hahn over Frankfurt am Main.
06	The four logs are plausibly compatible. They do not resolve which step caused the failure.

A Decision Chain — each step emitting a Receipt that references the prior step's Receipt — collapses the four-party reconciliation into a single replayable record. The Chain section that follows describes the construction.

The public-sector mirror

The same architecture appears in the public sector.

> DIIA.AI · UKRAINE.
>
> Ukraine's national AI assistant inside the Diia public-services platform has been operational since September 2025 and is issuing automated income certificates to citizens at scale (WEF, Making Agentic AI Work for Government, April 2026). When one of those certificates is challenged — wrong amount, wrong tax year, wrong identity — the citizen has the right to ask why. The agency has no per-decision artefact to answer with. The Gap is not a financial-services problem; it is the operating condition of every public-sector AI deployment running today.

The WEF flags fraud detection, eligibility, benefit calculation, grant allocation, and tender preparation as commercially attractive but ethically loaded — precisely the functions where a citizen has the right to ask why a claim was denied, and where the framework has no per-decision mechanism to answer. Its recommended mitigations collapse into the same familiar triplet — human-in-the-loop validation, audit trails, explainability — none engineered as a verifiable artefact.

Agentic AI is not a category where the existing toolkit is less effective. It is a category where the toolkit does not fit — it assumes a human reviewer who, at agent throughput, does not exist.

THE GAP

The artefact a regulator wants is the one nobody emits.

Why Existing Solutions Fail

Four candidates compete to close the Decision Boundary Problem. Each fails in a specific and instructive way — explainability tools, logs, governance frameworks, and model cards. The scenario below shows how all four arrive at the same dead end inside a single inspection.

Scenario · An inspection with four artefacts and no decision.

Each candidate fails the same way: it sits outside the moment of execution.

01 A regulator opens an inspection of a bank's adaptive-pricing engine.

02 With Decision Receipts, every disputed decision is retrievable, replayable, and verifiable against its policy snapshot — the inspection collapses into a query.

03 Without them, the bank produces SHAP values from the current model. The model has been retrained twice since.

04 It produces audit logs. The logs rotated after ninety days; the relevant window is gone.

05 It produces an ISO/IEC 42001 dossier. It describes the system, not the decision.

06 It produces a model card. Current as of last quarter, not last March.

07 Four governance artefacts. None is the decision.

How each candidate fails

Each of the four artefacts answers a different question, and each sits outside the moment of execution.

01 **Explainability tooling.** Answers "why did the model produce this output?" by reconstructing behaviour after the fact. The reconstruction is an estimate, generated by tools whose own assumptions are open to question — and the firm controls when, how, and against which version of the model the explanation is run.

02 **Logs and monitoring.** Answer "what activity occurred?" The trail thins out as moving parts grow — under retraining, vendor opacity, and LLM-mediated reporting. Logs are unstructured, retention-capped, and ops-editable.

03 **Governance frameworks.** ISO/IEC 42001, NIST AI RMF, OECD, EU AI Act procedural obligations — answer "is the firm responsible?" They sit above the decision and rely on the firm to flow requirements down. The order-of-magnitude AICDI gap between claim and evidence is what that flow-down looks like in practice.

04 **Model cards and conformity dossiers.** Answer "what is this artefact, in the abstract?" They describe the model at documentation time. They are not bound to any decision the model produced. A firm can hold a perfect ISO/IEC 42001 certificate and still be unable to prove this loan denial, on this date, for this applicant.

Four artefacts. No decision.

Logs · What happened
Explainability · Why the model behaved
Frameworks · Whether the firm is compliant
Model cards · What the model is

CANNOT PROVE THIS DECISION

Decision Receipt

What was decided
By what
Against which rules

PROVES THIS DECISION

MeshQu FOUR ARTEFACTS. NONE IS THE DECISION.

Only one produces proof at the moment it matters.

The Missing Primitive

The artefact that closes the Gap is structurally simple. A **Decision Receipt** is a record, emitted by a regulated AI system at the Decision Boundary, that binds the input, the policy in force, the model, and the output into a single signed object.

This is not a governance layer. It is an infrastructure requirement. Systems that cannot produce this artefact are likely to struggle under regulatory or legal scrutiny in high-stakes decisions.

What the Receipt contains

A Decision Receipt records:

A hash of the inputs the system saw.

A reference to the policy version in force at that moment.

The identity and version of the model or agent that produced the output.

Any human-oversight signal — review, override, attestation — attached to the decision.

The output the system produced.

A timestamp.

A cryptographic signature.

SYSTEM OUTPUT

Schema	meshqu-receipt@v1
Issuer	org:lender-bank-uk
Decision ID	meshqu:credit:7f3a-9c1e
Policy Snapshot	credit-risk@v3.1.4
Inputs	hash:sha256:8b41...e2d9
Model	lender-llm@v2.6.0
Outcome	DECLINED
Oversight	human-reviewed
Timestamp	2026-03-12T14:02:18Z
Integrity	sha256:9c1e...4d8f
Signature	ed25519:9b...f3
Anchor	rekor-log:idx-...

Four properties that distinguish the Receipt

01 **Created at execution.** The Receipt is emitted by the decision itself, not reconstructed afterwards. There is no window for the firm's narrative to drift from the record.

02 **Cryptographically bound.** The signature ties the Receipt to a specific signing key. The record is tamper-evident: any modification invalidates the signature.

03 **Replayable.** The input hash and policy reference let any party reproduce the conditions of the decision under the determinism parameters the Receipt records. The Receipt captures the bindings required to perform replay independently of the system's current state; what is reproducible is the decision against the policy in force, not necessarily every probabilistic intermediate computation.

04 **Independently verifiable.** Verification needs only a public key and a hash function. No outside party — regulator, claimant, auditor, counterparty — has to trust the firm's tooling. The signature proves the Receipt was emitted by the holder of a specific key; trust in the key itself comes from external anchoring — transparency logs, third-party signing services, and (where adopted) regulator-held verification keys. Verification does not require reliance on MeshQu as a central authority. Receipts can be signed using organisation-controlled keys and anchored to independent infrastructure, allowing verification without reliance on a single vendor. The primitive separates the cryptographic property from the institutional one, so each can be addressed on its own terms.

The artefact must also be portable. A regulator, auditor, or counterparty must be able to independently replay and verify the decision without requiring access to the originating system, production database, or vendor infrastructure. Verification must survive organisational boundaries, infrastructure changes, and vendor exit.

> NOT NEW GOVERNANCE — OPERATIONALISED GOVERNANCE.
>
> The EU AI Act's "traceability and reproducibility," AFM's "fully accountable regardless of technological complexity," the FCA's per-customer Consumer Duty outcome standard, and TRF/UNESCO's call for "verifiable practice" all point at the same artefact-shaped hole. The Decision Receipt is designed to fit it.

What the Receipt is not

The Decision Receipt sits in a family of signed-record primitives developed for adjacent problems. Naming the neighbours is the cleanest way to say what it is.

W3C Verifiable Credentials prove that an issuer made a claim about a subject. They are designed for identity attestations — a university issuing a degree, a regulator issuing a licence. They are not bound to a runtime decision and carry no policy-version or input-hash semantics.

Sigstore, in-toto, and SLSA attestations prove the provenance of a software artefact through its build pipeline. They answer "was this binary built from this source by this builder." They do not extend to the artefact's runtime behaviour.

EU AI Act Article 12 logging mandates automatic event recording for high-risk systems. It specifies what must be captured but not the cryptographic form, the binding to policy version, or the verifiability properties. Article 12 names the requirement; it does not standardise the artefact.

The Decision Receipt borrows from all three — the issuer-claim shape of Verifiable Credentials, the cryptographic anchoring of Sigstore, the regulatory mandate of Article 12 — and adds the property none of them carry alone: a signed binding of policy, input, model, and output, emitted at the moment a regulated decision is made, replayable against a frozen snapshot, verifiable by any outside party.

The Receipt sits in the same lineage as SBOMs and signed software attestations: machine-verifiable evidence that procurement, supervision, and audit are increasingly seeking, applied to the surface where decisions are made rather than where software is built.

The difference is not technical. It is about proof. The question changes from "do we believe this happened?" to "can we verify that it did?"

Where this sits

```
NOT

A governance framework
An observability layer
An explainability tool

SITS AT

The decision execution layer

INTEGRATES WITH

Decisioning systems
Policy engines
Workflow systems

EMITS

A verifiable artefact at the moment of decision
```

From Decisions to Decision Chains

Single decisions are the simple case. Most regulated AI workflows are multi-step: a credit application moves through KYC, sanctions screening, fraud scoring, affordability and pricing. A trade moves through pre-trade research, execution, and post-trade reconciliation. An agentic purchase moves through user mandate, agent search, merchant policy match, payment authorisation and fulfilment.

A single Receipt proves a decision. A Chain proves a system.

When each step emits a Decision Receipt that references the prior step's Receipt, the result is a **Decision Chain** — a verifiable, end-to-end record across systems and vendors, designed to remain replayable across jurisdictional boundaries where the cryptographic primitives are mutually recognised. The Chain inherits the Receipt's independence: any party can verify any segment without trusting any other party. A Chain is what replaces "trust the firm" with "verify the segment."

Fig. 3 shows a credit application moving through KYC, sanctions, fraud, analyst override, and affordability — every transition leaving a signed Receipt that references the prior one.

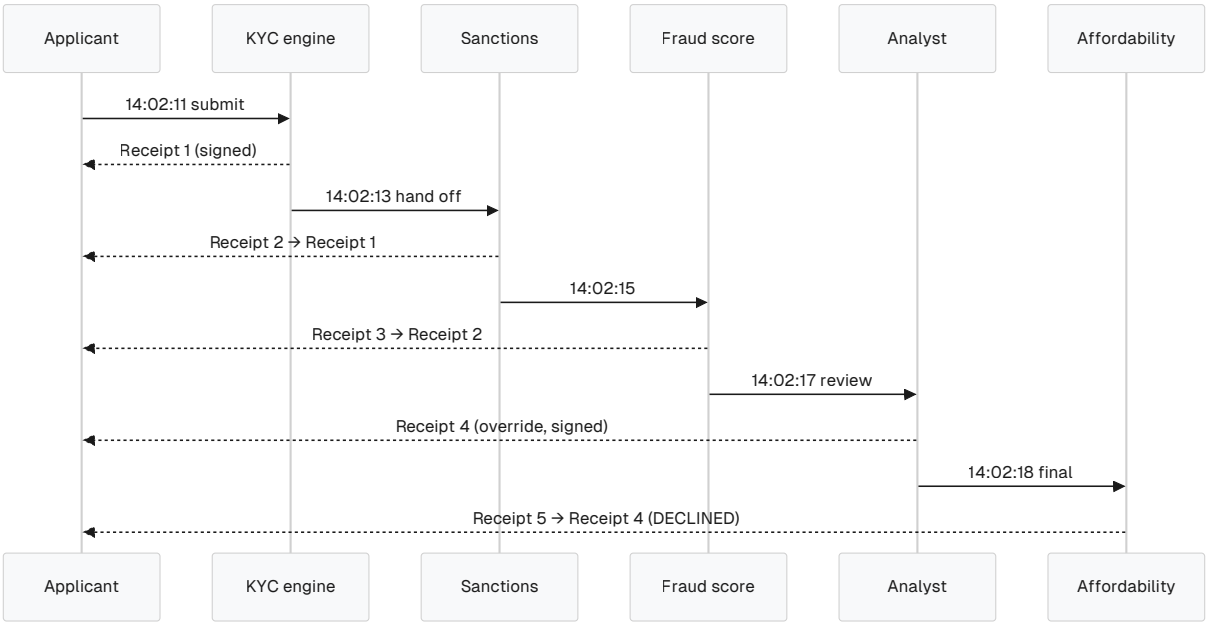


Fig. 3 — Each step emits a receipt referencing the prior step — the chain is the audit trail.

Eighteen months later, the chain is replayed end-to-end. Each step verifies against the next. The path is proven, not narrated.

What Chains solve that single Receipts do not

01 **Multi-actor accountability.** Accountability splits across actors — McKinsey's four-party case — and each party's contribution is non-repudiable.

02 **Vendor-boundary survival.** A model from a third-party provider leaves a chain segment that survives the vendor's later opacity.

03 **Human and machine review coexist.** A reviewer's attestation is itself a Receipt that joins the Chain, with the same verification properties.

For agentic systems, the Chain replaces the human-in-the-loop assumption with proof-in-the-loop: every action emits a Receipt that the policy was satisfied, and human review shifts from per-action gatekeeping to sampled oversight against verifiable evidence.

Consider a cross-border credit workflow. The originating institution signs the initial credit assessment. The risk-scoring provider signs its output. The executing institution signs the final approval. Each step produces a Receipt that references the prior step, forming a Chain that crosses jurisdictions, vendor boundaries, and institutional control without requiring any party to trust the others in aggregate. The integrity of the process emerges from the composition of independently signed decisions, not from any central authority.

Return to the four-party agent failure: the user, the agent platform, the merchant, the payment network. Today, four logs exist and none locates the decision. With a Chain, the user's mandate is a Receipt; the agent's merchant-policy match is a Receipt that references it; the merchant's fulfilment is a Receipt that references both; the payment network's settlement closes the Chain. The dispute resolves to a single replayable record that names the moment the agent committed to Hahn over Frankfurt am Main — and which party's policy permitted it.

This is the pattern MeshQu is designed to produce.

Scope and Limitations

Decision Receipts are an evidentiary primitive, not a control framework. They do not replace governance, model risk management, or human oversight. They make those instruments enforceable at the point they are tested. Four boundaries on where the primitive applies.

01 **Not every decision needs a Receipt.** Low-stakes, reversible, uncontested decisions — a recommendation engine surfacing a film, a search-result ranking — do not warrant the cost of signing and storage. The primitive is for decisions that can be challenged: credit, hiring, benefits, trades, agentic transactions, clinical triage, procurement awards.

02 **Latency and throughput floors exist.** A signing step adds a few milliseconds. For decisions made at sub-millisecond cadence — high-frequency trading microstructure, real-time ad auctions — the architecture has to be designed around batched or deferred signing. The primitive scales; it does not arrive free.

03 **The Receipt is itself sensitive.** A signed record of a medical triage or a benefit denial is regulated personal data. Storage, access control, and retention need to be engineered. Decision Receipts constrain logs by capturing the minimum required to prove a decision — they do not eliminate the privacy surface.

04 **The signature proves non-repudiation, not honesty.** A firm holding its own signing key can sign whatever it wants. Receipts become trustworthy through external anchoring (described above) and through the obligation to emit one for every decision in scope. Both are governance choices, not properties of the primitive.

The Receipt is necessary, not sufficient. It closes the artefact-shaped hole. The institutional question — who must emit, who holds keys, who can audit — sits above it.

Decision Receipts strengthen evidentiary integrity and replayability. They do not replace broader governance, operational controls, or institutional accountability obligations. They sit alongside those instruments as the layer that makes them testable.

Receipts establish what the system did against the policy in force. They do not establish that the policy was right, that the inputs were appropriate, or that the outcome was normatively correct. Correctness remains a separate question, addressed by policy validation and model risk management.

Implementation realism

Implementing Decision Receipts introduces operational overhead — signing latency, storage, and key management. The costs are measurable and bounded. Cryptographic operations complete within milliseconds and can run asynchronously or at the edge. Storage scales predictably with decision volume, and admits compression and selective retention.

Adoption is incremental. The receipt-emitting layer sits alongside existing decisioning systems, policy engines, and workflow orchestrators rather than replacing them — typically as an inline call or a sidecar that observes the decision at the moment it is committed. Policy snapshots, input hashes, and signing keys are integrated where the decision is made; downstream verification consumes the artefact, not the system. Existing infrastructure remains in place; the Receipt is added at the boundary.

The trade-off is explicit: a small increase in system overhead in exchange for the ability to produce verifiable evidence under scrutiny.

Practical Implications

The Receipt and the Chain reshape what each party in the AI accountability chain can do — supervisors, regulated firms, AI operators, and the people whose lives the systems decide. The pairs below show what changes for each.

PORTABLE VERIFICATION.

A Decision Receipt is not merely viewable. It is independently verifiable. Replay and verification can be performed from a self-contained evidence package containing the receipt, policy snapshot, integrity proofs, and transparency records. Verification does not require access to MeshQu infrastructure, production databases, or the original execution environment.

Supervision today samples systems. With Receipts, it interrogates decisions.

For regulators

- 79% of regulators want supervisory tools (CCAF Global AI Report, 2026)
- Receipts complement model-probing initiatives such as Project Noor
- Audits become queries: show me every decision in March 2026 made under policy v3.2 with override flag set
- Answers in seconds, without trusting reconstruction

For financial institutions

- Consumer Duty outcome demonstration → query
- RTS 6 (MiFID II algorithmic-trading) ex-post reconstruction → query
- Anti-money-laundering (AML) lookback under enforcement → query
- EU AI Act Article 12 record-keeping → query
- Compliance cost per inspected decision falls sharply

A note on timing. Core regulatory obligations are already coming into force; harmonised technical standards and implementation practice continue to mature. During this 2026–2027 window, architectures that produce verifiable decision-level evidence are likely to become the stronger practical compliance posture — not because standards mandate them, but because the requirements they satisfy are already binding.

The implications outside finance are sharper still — and the asymmetry between firms and the people they decide on collapses entirely once the artefact is shared.

For AI operators outside finance

- Public-sector deployments, agents, marketplaces, healthcare, legal
- McKinsey's "liability vacuum" exists because no party can demonstrate which step caused a failure
- With Decision Chains, any party can
- Operators with provable artefacts will be easier to underwrite
- Operators without them face growing scrutiny

For counterparties and end users

- The consumer challenging an AI-mediated decision
- The bidder challenging a procurement award
- The claimant disputing a benefit calculation
- Today: contest the firm's narrative with no independent artefact
- With Receipts: hold the same evidence the regulator holds, and the firm cannot edit

For vendors and platforms

- Compliance is becoming a procurement filter — clients increasingly ask whether vendor systems can produce tamper-evident, decision-level evidence
- Multi-tenant logging creates architectural tension between traceability and data minimisation; Receipts capture the minimum binding artefact without exposing raw inputs

Longer-horizon implications

- Vendor-boundary survival: when a model provider's service later changes or sunsets, the customer's prior decisions remain provable
- The Receipt becomes a standardised evidentiary object across clients, jurisdictions, and audit contexts

Scenario · A claimant disputes a public-sector decision.

With Receipts, the artefact is the same on both sides of the table.

01	A claimant disputes a denied benefit calculation made by an automated public-sector system.
02	Today, the claimant submits a complaint, waits eight weeks, and receives a narrative response written by the agency.
03	With Receipts, the claimant retrieves the Receipt of their own decision, alongside the policy version that ran it.
04	The claimant's lawyer verifies the signature locally. The Receipt either matches the agency's claim or it does not.
05	The bargaining position is no longer one-sided. The artefact is the same on both sides of the table.

The cost of compliance falls. The cost of not having proof rises.

Conclusion

AI governance fails not because the rules are wrong, the regulators are weak, or the firms are uncooperative. It fails because the system does not produce evidence at the Decision Boundary. Frameworks, model cards, sandboxes, explainability tools, and audit logs each address part of the problem. None addresses the part that matters when accountability is tested.

The Decision Receipt introduces a verifiable artefact at that boundary: a signed record, anchored to policy version, that any outside party can verify without trusting the firm. It does not replace governance — it makes governance enforceable at the point it is tested.

> TRF / UNESCO call for "verifiable practice." AFM, in AI in Capital Markets: Balancing Innovation and Integrity (April 2026), points toward continuous evidence and interoperable oversight as the longer-term direction for capital-markets supervision, rather than post-hoc inspection alone. EU AI Act Article 12 names the requirement: traceability and reconstructability of system behaviour. APRA's April 2026 letter to industry reaches the same point from prudential supervision. All converge on the same artefact-shape.

Without proof at execution, governance is narrative, audit is reconstruction, and liability is contested on the firm's terms. With proof, decisions become verifiable and queryable on the same evidence regulators already ask for.

Supervisors don't need more reports. They need to query decisions.

WITH PROOF

The decision stands on its own.

Authority that does not reach the decision is theatre.

About the Author

Sam Carter is the founder of MeshQu. He graduated with an MEng in Software Engineering in 2009 and has spent the seventeen years since building B2B software inside regulated industries — banking, insurance, fintech, trade finance, and logistics.

His recent work has concentrated on financial infrastructure and decisioning systems: trade-finance platforms (Marco Polo Network, Kanexa), Web3 financial infrastructure (OpenTrade), purchase-order financing (Tameed, Saudi Arabia), private-debt portfolio risk and reporting (Five Sigma), and rules-engine programmes where policy, approval, and execution have to be reconciled under audit. The recurring brief — across institutions, jurisdictions, and vendors — has been governance-heavy workflows in which the cost of a missing or reconstructed audit trail is operational and regulatory rather than theoretical.

MeshQu emerged from that work. Over the past eighteen months, building and researching AI inside regulated environments, the same wall kept appearing: frameworks describe what should happen; systems cannot prove what did. The Decision Proof Gap is the synthesis of that pattern. MeshQu is an attempt to operationalise an answer to it.

About MeshQu

MeshQu is Decision Assurance infrastructure. Most systems can explain decisions. MeshQu makes them provable. It enables systems to produce verifiable Decision Receipts at the moment a decision is made — binding inputs, policy, actor, and outcome into a tamper-evident record that can be replayed and independently verified. Designed for environments where decisions must withstand scrutiny: financial services, public-sector systems, and clinical decision support.

Use-case demonstrations across regulated workflows are operational. Each demo's Receipts can be independently verified without credentials and without involvement from MeshQu — the verification path is the proof. MeshQu is in early-stage discussions with design partners and regulators in financial services and the public sector.

For pilot enquiries, regulatory engagement, or research correspondence: contact@meshqu.com.

The goal is simple: move from reconstructed explanations to provable decisions.

References

Multilateral and regulatory primary sources are preferred; survey and consulting material is cited where it adds comparative breadth.

[01] **Cambridge Centre for Alternative Finance (CCAF), BIS, IMF, WEF, IDB, CGAP, Arab Monetary Fund (2026).** 2026 Global AI in Financial Services Report — Adoption, Impact and Risks. Survey of 628 organisations across 151 jurisdictions, late 2025–early 2026.

[02] **Thomson Reuters Foundation and UNESCO (2025/2026).** Responsible AI in Practice: Insights from the AI Company Data Initiative (AICDI). Survey of 2,972 companies.

[03] **Autoriteit Financiële Markten (2026).** AI in Capital Markets: Balancing Innovation and Integrity. AFM, April 2026.

[04] **European Union (2024).** Regulation (EU) 2024/1689 — Artificial Intelligence Act. Official Journal of the European Union. Articles 12, 17; Annex III.

[05] **Roberts, H., Taddeo, M., Floridi, L. (2026).** "The hybrid regime complex of AI governance." Global Policy.

[06] **World Economic Forum (2026).** Making Agentic AI Work for Government. April 2026.

[07] **Gartner (2026).** Forecast on agentic-AI project cancellations, cited in WEF (2026).

[08] **McKinsey · QuantumBlack (2025).** The Agentic Commerce Opportunity. October 2025.

[09] **Financial Conduct Authority (2022–23).** Consumer Duty (PS22/9). In force since 2023.

[10] **ISO / IEC (2023).** ISO/IEC 42001 — Artificial Intelligence Management System.

[11] **National Institute of Standards and Technology (2023).** AI Risk Management Framework (AI RMF 1.0).

[12] **OECD (2019, rev. 2024).** Recommendation of the Council on Artificial Intelligence.

[13] **Council of Europe (2024).** Framework Convention on AI, Human Rights, Democracy and the Rule of Law.

[14] **UNESCO (2021).** Recommendation on the Ethics of Artificial Intelligence.

[15] **Bank for International Settlements — BIS Innovation Hub (2025–26).** Project Noor — supervisory model-probing initiative for AI in regulated environments.

[16] **AI & Partners (2026).** UK AI Assurance Market Analysis · Insight 6. April 2026.

[17] **ESMA (2017–).** Regulatory Technical Standard 6 (RTS 6) — Algorithmic Trading Record-Keeping. Under MiFID II.

[18] **Anthropic (2025).** Research note on reward-hacking and monitoring evasion, cited in AFM (2026).

[19] **Government of Ukraine (2025–).** Diia.AI — public-services AI assistant. Operational since September 2025.

[20] **Lundberg, S., Lee, S.-I. (2017).** A Unified Approach to Interpreting Model Predictions (SHAP). NeurIPS.

[21] **Ribeiro, M., Singh, S., Guestrin, C. (2016).** "Why Should I Trust You?" Explaining the Predictions of Any Classifier (LIME). KDD.

[22] **Sundararajan, M., Taly, A., Yan, Q. (2017).** Axiomatic Attribution for Deep Networks (Integrated Gradients). ICML.

[23] **Australian Prudential Regulation Authority (2026).** Letter to Industry on Artificial Intelligence — Summary of Common Weaknesses and Expectations for Regulated Entities. APRA, April 2026.

[24] **Nannini, L., Smith, A. L., Maggini, M. J., Panai, E., Feliciano, S., Tiulkanov, A., Maran, E., Gealy, J., & Bisconti, P. (2026).** AI Agents Under EU Law. arXiv:2604.04604.

— THESIS

Authority that does not reach the decision is theatre.

CONTACT

MeshQu Ltd
Author · Sam Carter
contact@meshqu.com

DOCUMENT

WP–PROOF–01
v1.0 · May 2026

CLASSIFICATION

Public Distribution
Recipient: N/A

RECOMMENDED CITATION

Carter, S. (2026). The Decision Proof Gap.
MeshQu White Paper WP-PROOF-01. Zenodo.
https://doi.org/10.5281/zenodo.20055736

VERSION	PUBLISHED	DOI
v1.0	May 2026	10.5281/zenodo.20055736