# The Agent Accountability Gap

## An agent workflow can report success without independent evidence that the business outcome was legitimate.

AI agents now pay vendors, change master data, post to ledgers and move cash, often across several platforms and a chain of delegated authority. When one of those runs is questioned by an auditor, a risk committee, a regulator or the board, the main record of what happened usually comes from the platform that ran it. This briefing sets out the gap and what to do about it.

### 1 · THE CONSEQUENTIAL WORKFLOW

> **HYPOTHETICAL EXAMPLE: INVENTED COMPANY, POLICIES AND AMOUNTS**
> A large enterprise runs four agents on two ERP-connected platforms: an **intake agent** receives a supplier's request to change bank details; a **vendor-master agent** updates the ERP supplier record; an **AP agent** schedules payment of open invoices; a **treasury agent** releases the batch to the bank. Mid-run, the Vendor Master Policy is reissued with a new five-day hold on bank-detail changes, and the AP manager's delegation is reassigned. The run ends: *"Batch released, 14 payments, status: success."*

### 2 · WHAT EXISTING SYSTEMS CAN ESTABLISH, AND WHAT THEY CAN'T

| Layer | Can establish | Doesn't establish on its own |
|---|---|---|
| Governance | That policies, owners and approved uses exist | What a specific action did |
| Identity & access | Which agent identity acted, with which scopes | Whether the business mandate survived each handoff |
| Runtime control / gateway | That each call was allowed at the control point | The downstream result; whether the whole run stayed in scope |
| Observability | Traces, tool calls, timings | Authority; independence from the acting system |
| Agent platform | Its own account of what it did | Anything independent of itself |

In the example, every layer reports green. Three questions remain open:

**Stale policy at execution**
Was each action judged against the policy version in force *when it occurred*, or against a cached or since-overwritten version?

**Authority across agents**
Each agent stayed within its own permissions. Did the *run as a whole* stay within the mandate that was originally delegated?

**Reported vs. reconciled**
"Success" is the agent's assertion. Do the ERP and the bank show the outcome the agent reported?

### 3 · WHY INDEPENDENT VERIFICATION MATTERS

**The system that performs the action should not vouch for the action.** That's a statement about where a record is produced, not a judgment about its accuracy. A record produced inside the system under evaluation is a non-independent witness: it can be accurate, but on its own it can't establish independent assurance. It's the same logic as segregation of duties. COSO's 2026 guidance on generative AI (a publication, not a requirement) makes the parallel point: *"Controls should treat outputs as claims requiring validation, rather than as facts to accept by default."* In a multi-agent workflow there's a practical problem too. Each platform logs only its own slice, so nobody holds the mandate and the full trajectory together.

## 4 · WHAT INDEPENDENT EVIDENCE FOR A CONSEQUENTIAL RUN NEEDS (DESIGN PRINCIPLES)

These are design principles. The hosted demo shows binding each finding to the policy version in force, and keeping policy and mandate results separate, on evaluator-produced synthetic findings. Reconciliation and INDETERMINATE results appear there only in generated background data. Reconciling downstream outcomes across several records is in development.

1. **The consequential action as the unit of evidence**, with the run as its context.
2. **Historical rule binding:** each action tied to the policy version in force when it occurred (delegation versions are in development). Today's policy must not rewrite yesterday's evidence.
3. **A trajectory-level mandate test:** "policy passed" and "stayed within mandate" are separate results. Permitted steps can add up to an unauthorized result. (The term *authorization continuity* comes from Zhang & Zhang, arXiv:2607.23586.)
4. **System-of-record reconciliation:** the agent's report and what a registered system of record independently reports, kept apart and compared, with differences recorded as evidence and never silently resolved.
5. **Reference-bounded findings:** where no authoritative rule applies, the result is indeterminate, not an invented "should."

## 5 · A SYNTHETIC DEMONSTRATION

**HOSTED DEMO · FICTIONAL COMPANY · SYNTHETIC DATA · NOT A CUSTOMER DEPLOYMENT**

Weaveframe's hosted demo is a read-only review environment for a fictional "Weaveframe Demo Enterprise" with approximately 50 synthetic agents across multiple workflows. In Weaveframe's hosted demo, a synthetic environment, a small set of featured findings was produced by Weaveframe's evaluator against synthetic approval and authority records. In them: an agent that reports "approved" with no independent approval record gets **MISSING_EVIDENCE**, not PASS. An approval from the acting system itself also gets MISSING_EVIDENCE. A $200,000 action under a $100,000 mandate, independently approved, is still a **VIOLATION**, because an approval doesn't expand a mandate. Each featured finding is bound to the policy version in force when the action occurred, and the record shows that policy's name, version, hash and effective window. Findings use five states: PASS, VIOLATION, MISSING_EVIDENCE, EXCEPTION, INDETERMINATE.

## 6 · WHERE REGULATION STANDS, STATED CAREFULLY

No regulator or standards body we have reviewed requires independent evidence for each consequential agent action. The direction of travel is toward logging, traceability, human oversight and independent verification, mostly at system and organizational level. The **EU AI Act** requires high-risk systems, as it defines them, to allow automatic event logging (Art. 12) and to be designed for human oversight (Art. 14), applying from Dec 2027 (Annex III) and Aug 2028 (Annex I). **NIST** launched an AI Agent Standards Initiative on Feb 17, 2026, and its NCCoE draft concept paper sought input on agent identity, authorization, auditing and non-repudiation. **Singapore's** agentic-AI framework (IMDA, Jan 2026; guidance) says deploying organizations and the humans overseeing agents remain accountable. **COSO** (Feb 2026; guidance) applies its internal-control framework to generative AI, including agents performing multi-step tasks.

## 7 · WHAT TO DO NEXT: FIVE QUESTIONS FOR YOUR TEAM THIS QUARTER

1. Which agents can move money, change master data, post to the ledger or file on our behalf, including handoffs across platforms?
2. For one of those workflows, which policy version and delegation applied to each consequential action *at the time*?
3. Did the run as a whole stay inside the mandate that was granted, or only each step?
4. Does the system of record match what the agent reported, checked through a read path the agent doesn't control?
5. Which of those answers rests only on the acting agent's own account?

### Next step: a free Agent Execution Assurance Review

Bring one consequential finance, procurement or ERP-connected workflow to a free, 30-minute, founder-led conversation. Nothing to install, and no production data, customer records, credentials or NDA needed. Afterwards you get a concise summary: the workflow discussed, the key assurance questions, and recommended next steps.

**Request your free Review: email jon@weaveframe.ai or reply to the message that brought you here · Jon Odalen, Founder**

Weaveframe builds independent assurance for consequential enterprise AI-agent actions. It has a working prototype and is actively pursuing its first design partners. No production deployment is described here. Weaveframe does not block, stop or govern agents and does not issue audit opinions, certifications or legal attestations. Scenarios are hypothetical or synthetic, as labelled. Sources: Weaveframe working paper *Independent Agentic Assurance* (v11); COSO, *Achieving Effective Internal Control Over Generative AI*, Feb 23 2026; Regulation (EU) 2024/1689, Arts. 12, 14, 113; NIST CAISI, AI Agent Standards Initiative, Feb 17 2026; NIST NCCoE, Software and AI Agent Identity and Authorization concept paper (draft), Feb 5 2026; IMDA Singapore, Model AI Governance Framework for Agentic AI, 2026; Zhang & Zhang, arXiv:2607.23586.