Argentis Labs

Working Paper 2.5 · 2026

The approval click is now the evidence

the click now decides who is liable — and only an operator-independent record can be verified, not believed

Argentis Labs Research

2026-06-14

Between March 5 and April 29 of this year, three global jurisdictions moved the same way: toward pinning accountability on named individuals. First, the Federal Court of Australia found Star Entertainment's former CEO and its Chief Legal and Risk Officer personally liable under section 180 of the Corporations Act for failing to oversee money-laundering risk: the non-executive directors were cleared; the two operating officers the court ruled on were not. Then on March 10, the US Department of Justice issued its first department-wide Corporate Enforcement and Voluntary Self-Disclosure Policy and built its corporate-leniency framework around the pursuit of individual wrongdoers. Finally, in late April the UK's Crime and Policing Act 2026 received Royal Assent; section 250 takes effect on June 29 and attributes any criminal offence committed by a senior manager, acting within actual or apparent authority, to the corporation itself.

Individual liability is old; the machinery is new: enforcement that prices leniency in named individuals, attribution that runs through operational roles, systems that record, at click granularity, who approved what. The migration is evidentiary. For the enterprise deploying agentic AI, the unit of that evidence, increasingly the unit of accountability itself, is the approval: rendered against an output the approver did not produce, by a system the approver does not operate, in a log the operator can rewrite.

That record carries an asymmetry worth stating plainly. The current substrate is good enough to convict you and not good enough to clear you.

Attribution just stopped requiring the boardroom

The UK change is the most visible because it has a date. Section 250 repeals and replaces the senior-manager attribution regime of sections 196–198 of the Economic Crime and Corporate Transparency Act 2023, which confined attribution to a schedule of economic crimes. The new provision extends the same model to every criminal offence on the books. "Senior manager" is defined functionally: anyone who plays a significant role in deciding how a substantial part of the organisation's activities are managed, or in actually managing them. The test reaches below the board, and in a decentralised enterprise it reaches well below it.

US federal law has attributed employee crimes to the corporation since *New York Central & Hud-*

son River Railroad v. United States in 1909. Respondeat superior reaches any employee acting within the scope of employment for the corporation's benefit, a broader rule than the one the UK just adopted. Section 250 is the rest of the common-law world arriving where US executives have lived for a century.

The American development is not the attribution rule. It is the mechanism. The Department of Justice's March 10 policy offers a declination (no prosecution of the company) in exchange for voluntary self-disclosure, full cooperation, and remediation. Full cooperation means producing all non-privileged facts relevant to the misconduct and identifying every individual involved, whatever their position, status, or seniority. Nine days later came the first resolution under it: a declination for medical-device maker Balt SAS, announced alongside the indictment of a Balt executive and a consultant for the underlying bribery scheme, an indictment a grand jury had returned on March 4. The company was spared; the individuals were not.

Read together, the structure is clear: the company's path to leniency runs through its own records, and the trail identifying who approved what is increasingly part of the consideration tendered for the declination. The individual doctrine was already old: *Dotterweich* (1943) and *Park* (1975) made a responsible relation to the violation, plus the power to prevent it, a substitute for personal knowledge. And recent practice extended it down the org chart: a criminal conviction for Uber's former chief security officer, affirmed by the Ninth Circuit in 2025, and a $250,000 personal penalty for MoneyGram's former chief compliance officer. The SEC's 2023 action against SolarWinds' CISO (charged personally over the company's security disclosures, dismissed in late 2025) changed the doctrine less than it changed the targeting: attribution now begins at the officer level.

Australia supplies the parallel rather than the exception. The Star judgment turned on section 180's duty of care applied to oversight of non-financial risk, the same theory Delaware courts run under *Caremark*, sharpened by *Marchand v. Barnhill* into a duty to monitor mission-critical risk, and priced by the Boeing 737 MAX derivative settlement at $237.5 million. APRA's CPS 230 makes the board ultimately accountable for operational risk, third-party providers included, and AI sits within that perimeter, the same posture US banking examiners have taken toward models since SR 11-7 in 2011, carried into its April 2026 successor, SR 26-2. No Delaware court has yet heard an AI oversight claim. Everything above suggests the first one will arrive with its analysis already supplied.

The approval attaches the liability

Human-in-the-loop was adopted as a liability answer: a person reviews the output, and the person's judgment confers legitimacy. Under the regime described above, the review does something else. It attributes.

Consider the standard failure. An agent executes an action permitted by its tooling but violating a constraint never encoded: a conflict rule, a jurisdictional restriction, a client instruction held in someone's head. A reviewer approves the output; the breach surfaces months later. Section 250 makes what follows precise. Attribution runs through a senior manager's offence (an agent acting alone creates no corporate offence by this route), and the offence, not the click, is the statutory element. The click is where the workflow becomes legally legible: in an AI-mediated workflow, the approval is often the only human act available to anchor conduct, knowledge, or recklessness. "The system guided me" is adoption, not absolution: the output became the

manager's act at the click, and reliance on a system the manager could not inspect reads less like diligence than recklessness. The EU AI Act's Article 14 states the same demand in regulatory terms: oversight that is effective rather than present. Whether a given review was meaningful or merely ceremonial is a factual question, and it is answered from the record.

The gap suggests a perverse move: remove the approver, run the agent autonomously, and the section 250 route closes. No senior-manager offence, no attribution. The shield is narrower than it reads. Mens rea migrates to the deployment decision, since the law has long attributed an instrument's conduct to whoever set it in motion. And the officer-liability provisions that pattern UK regulatory law run opposite to section 250: where a strict-liability corporate offence is attributable to an officer's consent, connivance, or neglect, the officer is personally guilty, and the absence of oversight is the neglect. Article 14, CPS 230, and *Caremark*'s first prong reach the same place: no oversight is not a defence but the easiest governance failure to plead.

Now place the approval record inside the DOJ mechanism. The organisation discovers the breach, weighs the declination market, and discloses. Cooperation means producing the approval trail: authenticated by the producing party, attributed to named individuals. The approver's click, generated by the operator's own system, arrives in the prosecutor's file as an admission with a timestamp.

Whose interests align at that moment is worth tracing. In the US, the company's leniency scales with the specificity of the individuals it identifies. Under section 250, its defence runs the other way, toward showing that no senior manager committed an offence at all, leaving the approver alone in the frame if prosecutors disagree. Both fights are fought on a record the company operates. And neither the declination nor the disavowal protects the person named inside it.

This is where the asymmetry does its work, and it belongs to evidence law before it belongs to technology. Admissibility is not the issue; weight is. Offered against the producer, the log is an admission and carries weight naturally. Offered in the producer's favour, it is self-serving attestation asking the trier of fact to take the word of the people whose conduct is at issue about what the system presented, what alternatives were suppressed, what the trail contained before cooperation became the strategy. The record convicts on its face and clears nothing beyond it. And the same record that fails under adversarial scrutiny before a court fails under underwriting scrutiny before a D&O insurer, who asks the question years before any prosecutor does.

Observability and evidence are different planes

The reflex response is instrumentation: more tracing, more spans, more dashboards. The enterprise AI observability stack (Datadog, LangSmith, Langfuse, and their peers) is good at what it is built for, which is engineering. Traces exist to debug systems; they live under retention policies, writable by the operator, keyed to the operator's identity. None of that is a defect. It is the design of the engineering plane.

The failure is the conflation of planes. An organisation presenting observability data as its evidentiary record exhibits the failure mode of *Locally Correct, Globally Wrong*: every log line is well-formed (locally correct) while the relational property the situation demands, independent verifiability, structurally fails because the operator holds write access and the keys (globally wrong). A million perfect spans attested by the defendant are, to a prosecutor, a million statements by an interested party.

What the regulatory plane demands is a record whose integrity depends on no one's promise, least of all the party whose conduct is in question. Telemetry and evidence are different infrastructure categories with different threat models, not different configurations of the same one. Observability cannot solve evidentiary reconstruction because it was never asked to; an architecture that produces only the first has produced half of what the enterprise needs.

Verification cannot require the operator's cooperation

The second artifact is specified in our Working Paper 02, *Daubert Meets ChatGPT*. The substrate it defines has four properties, each answering a weakness in the operator-attested model. The system that produces the work and the system that records it run under separated identities, so the record's integrity never depends on any party whose interests may later diverge from it, the operator's included. The storage layer refuses deletion and alteration physically, beneath any account or operator instruction, so the integrity claim rests on the medium rather than on policy. Records carry public-key signatures, so a regulator, an opposing party, or a court verifies them without the operator's cooperation. And each matter's record is isolated, so producing the trail for one regulated artifact exposes nothing else.

A substrate with those properties inverts the asymmetry. The record remains available to a prosecutor: nothing about immutability shields misconduct, and a tamper-evident trail is often worse for a wrongdoer than a mutable one. What changes is the exculpatory side. The approver who acted correctly can demonstrate what occurred: what the system presented, what controls were enforced, and what the record contained before any party had an incentive to revisit it. Those facts no longer depend on the operator's testimony or the employer's assurances. They stand or fall on independent verification. The question ceases to be whether the record is trusted and becomes whether the record can be verified.

The fifth question is already drafted

The voir dire of an AI-assisted workflow ends, sooner or later, at a single question: how can the trier of fact verify that the workflow you describe is the workflow that actually occurred? After June 29, the alignment is complete. Across the common-law world, the person answering that question is, with increasing probability, an individual: a senior manager within the meaning of a statute, named in a record their employer produced in exchange for its own leniency.

That makes the problem less about AI and more about proof. The systems making decisions are becoming more autonomous. The people accountable for those decisions are not. As execution moves to software and accountability remains human, independent reconstruction is the bridge between them — what lets the human answer be tested rather than merely trusted. There are only two ways to cross it. One is a promise from the operator of the system: this is what happened, these were the controls, this was the record. The other is a record whose integrity does not depend on the operator at all: verifiable independently, against a key no party to the dispute controls.

The distinction is not technical. It is evidentiary. One asks the trier of fact to trust the witness; the other lets it verify the claim. The first wave of enterprise AI sold answers. The next wave will be judged by whether those answers can be reconstructed, attributed, and independently verified. As autonomy increases, that capability stops being a compliance feature and becomes part of the deployment decision itself: a requirement the regulation surfaced rather than imposed.

References

Crime and Policing Act 2026, s.250 (corporate criminal liability). legislation.gov.uk

New York Central & Hudson River Railroad Co. v. United States, 212 U.S. 481 (1909). justia.com

United States v. Dotterweich, 320 U.S. 277 (1943). justia.com

United States v. Park, 421 U.S. 658 (1975). justia.com

United States v. Sullivan (Uber CSO), No. 23-927 (9th Cir. Mar. 13, 2025). uscourts.gov

In re Thomas E. Haider (MoneyGram CCO) — FinCEN settlement release. fincen.gov

Marchand v. Barnhill, 212 A.3d 805 (Del. 2019). courts.delaware.gov

In re The Boeing Co. Derivative Litigation — Stipulation of Settlement (Del. Ch. 2021). osc.state.ny.us

U.S. Department of Justice (March 10, 2026). *Corporate Enforcement & Voluntary Self-Disclosure Policy.* justice.gov

U.S. Department of Justice (March 19, 2026). *Balt SAS Foreign-Bribery Resolution.* justice.gov

U.S. Department of Justice (March 10, 2026). *Corporate Enforcement & Voluntary Self-Disclosure Policy* — full policy text (PDF). justice.gov

ASIC v The Star Entertainment Group [2026] FCA 196 (Fed. Ct. Aus., March 5, 2026). fedcourt.gov.au

Federal Reserve (April 17, 2026). *SR 26-2: Revised Guidance on Model Risk Management* — cover letter (supersedes SR 11-7). federalreserve.gov

Federal Reserve (April 17, 2026). *SR 26-2: Revised Guidance on Model Risk Management* — attachment (full text, PDF). federalreserve.gov

Australian Prudential Regulation Authority. *Prudential Standard CPS 230 — Operational Risk Management.* apra.gov.au

Crime and Policing Act 2026 — gov.uk collection (Royal Assent April 29, 2026). gov.uk

Crime and Policing Act 2026, s.255 — commencement provisions. legislation.gov.uk

SEC v. SolarWinds Corp. & Timothy G. Brown — Litigation Release No. 26423, No. 1:23-cv-09518-PAE (S.D.N.Y. Nov. 20, 2025). sec.gov

EU AI Act (Regulation (EU) 2024/1689), Article 14 — Human oversight. artificialintelligenceact.eu