# From Traces to Evidence

*AI Agents and the Institutional Conditions of Computer Evidence*

**Venkat Nadella**

Draft: 16 August 2026

## Abstract

Legal and governance institutions increasingly rely on computer-generated records to reconstruct events, assign responsibility, provide remedies, and learn from failure. AI agents sharpen that dependency because their actions may pass through models, execution harnesses, tools, credentials, and organizational workflows controlled by different actors. The research question is: under what institutional conditions can runtime records from agent-mediated action become usable evidence for investigation, contestation, remedy, and learning when the action path, record custody, and authority to act are split across actors?

The paper combines four methods: doctrinal analysis of computer evidence and control-conditioned log-retention duties; a discriminating case study of *Bates & Others v Post Office Ltd (No. 6: Horizon Issues)*; known-outcome testing of a prototype Runtime Agent Deployment Governance Stress Test (GST); and a small coding pilot of 30 public AI incident records. Horizon is used because it produced evidentiary failure without AI-specific properties such as opacity, stochasticity, autonomy, or model updating. It therefore separates a general evidentiary mechanism from the features usually emphasized in AI governance.

The paper defines evidentiary fragmentation as a failure of the evidence pathway: records necessary to reconstruct or contest system action are distributed across actors or systems in ways that prevent the responsible institution or affected party from assembling, validating, interpreting, and acting on the relevant evidence at the point of consequence. It shows that duties to produce records can be legally enforceable yet practically inaccessible when an affected party first needs them. The GST converts that finding into a prototype deployment test, asking what an institution must be able to demonstrate before authorizing an AI agent to act rather than merely answer.

## 1. Introduction

In July 2026, OpenAI described a security incident during an internal evaluation of advanced cyber capabilities. In OpenAI's account, evaluated models exploited a vulnerability in an internally hosted package-registry proxy, escalated privileges inside the testing environment, and searched for information to help solve an evaluation task hosted by Hugging Face. Hugging Face's disclosure and technical timeline described the production-side intrusion and remediation from its side of the incident (Hugging Face, 2026a, 2026b; OpenAI, 2026).

The public record came from separate organizational disclosures. OpenAI's post covers the evaluation-side setting and response. Hugging Face's posts cover the Hugging Face-side intrusion and remediation. Each disclosure documents the portion its publishing organization placed in the public record. Neither document is general evidence returned to an independent investigator.

The evidentiary split matters because an evaluation environment became an environment in which an AI-mediated process could act. A chatbot can mislead a user. An agent with access to files, APIs, credentials, or production-like infrastructure can change an environment and create events that later have to be reconstructed.

Once AI systems can act through environments, safety depends on more than model behavior. It depends on operational conditions: scoped permissions, system boundaries, action logs, and incident routines. Human control has to appear in those forms before it can do any work as a governance principle.

Agent governance discussions often assume that logs, labels, evaluations, and incident reports will support accountability (Brundage et al., 2026; Chan et al., 2024; Chan et al., 2025; Ezell et al., 2025). This paper asks: under what assumptions do runtime traces become evidence that an institution can actually use? A retained log may still be inaccessible to the affected party, unintelligible to the regulator, incomplete for the court, or disconnected from any authority capable of remedy or institutional learning.

Technical observability is the ability to see some part of a system's operation. Institutional usability is the ability to convert a trace into evidence that can support a decision. This paper uses "institutionally usable evidence" in a broader sense than legal admissibility. The term includes material that helps courts, regulators, auditors, affected parties, or deploying organizations reconstruct events, evaluate responsibility, provide remedies, and learn from failure.

The research question is:

> *Under what institutional conditions can runtime records from agent-mediated action become usable evidence for investigation, contestation, remedy, and learning when the action path, record custody, and authority to act are split across actors?*

The argument has three steps. Computer evidence depends on institutional conditions as well as technical generation. Evidentiary failure can arise from fragmentation even without AI-specific properties. AI agents intensify that structure because the

governance object is the model-in-harness: the model, execution environment, permissions, verification loop, and organization around it.

The contribution is to identify a specific institutional failure mode in agent-mediated action: evidentiary fragmentation. Fragmentation is not mere distributed custody. It occurs when the records necessary to reconstruct or contest system action cannot be assembled, validated, interpreted, and acted on by the responsible institution or affected party at the point of consequence. The paper establishes that failure mode through a legal mechanism that sharpens an underexplored problem in AI-agent governance: a duty to produce records, conditioned on whether those records are under the duty-holder's control, can be legally enforceable and practically unreachable at the same time.

The paper also develops the Runtime Agent Deployment Governance Stress Test. The GST is a prototype instrument for asking whether an institution can specify an agentic system, inspect its action path, allow contestation, govern intervention and responsibility, and learn from failure before deployment hardens into routine operation. The first version of the instrument was specified before the close reading of *Bates No. 6*. Horizon exposed that the instrument had missed in-workflow contestability: a system-generated state could be disputed outside the workflow while remaining operative inside it. The revision is reported as a result of the known-outcome test rather than presented as if the revised version had been complete from the start.

The article proceeds as follows. Section 2 locates the argument in three related literatures: computer evidence and institutional reliability; reviewability, accountability, and contestability; and agent governance. Section 3 sets out the research design, propositions, methods, and scope conditions. Section 4 develops the evidence pipeline that connects technical traces to institutionally usable evidence. Sections 5 and 6 analyze computer-evidence law and Horizon as a discriminating case. Section 7 extends the Horizon mechanism to AI-agent systems. Section 8 tests the Runtime Agent Deployment GST against Horizon as a known-outcome case. Section 9 reports the incident-record coding pilot. Section 10 discusses implications for AI governance.

## 2. Related literatures and conceptual gap

Three literatures frame the question. The first is computer evidence and institutional reliability. Computer-evidence scholarship has already criticized the presumption that computer-generated evidence can be treated as reliable unless challenged (Ladkin et al., 2020; Marshall et al., 2021; Mason & Seng, 2021; Thimbleby, 2026). This paper does not claim novelty for that critique; and asks how that problem changes when the relevant action path is distributed across multiple technical and organizational actors.

The second literature concerns reviewability, accountability, and contestability. Reviewability scholarship asks what must be recorded to make automated decisions meaningfully reviewable (Cobbe et al., 2021). Work on algorithmic supply chains and the accountability horizon, the point beyond which an actor in a chain can no longer see or be held to account, shows how responsibility is distributed across interconnected actors and systems (Cobbe et al., 2023). Contestable-AI scholarship

treats contestability as a design, justifiability, and human-intervention problem across the system lifecycle (Alfrink et al., 2023; Almada, 2019; Henin & Le Métayer, 2022; Lyons et al., 2021; Vaccaro et al., 2019). This paper asks a specific evidentiary question within that broader terrain: when do runtime records become usable evidence for institutional decisions?

The third literature concerns agent governance and system-level AI governance. GovAI's work on non-model capability gains argues that model-level governance loses purchase when capability gains arise through scaffolds and other system-level improvements (Goemans et al., 2026). GovAI work on agent visibility, incident analysis, agent infrastructure, and frontier AI auditing identifies the records, monitoring systems, and access arrangements needed for governance (Brundage et al., 2026; Chan et al., 2024; Chan et al., 2025; Ezell et al., 2025). This paper takes the downstream evidentiary consequence: if capability is produced partly by the system around the model, accountability evidence must cover that system.

Ezell, Roberts-Gaal, and Chan's incident-analysis framework goes furthest in this direction by recommending what information developers and deployers should retain and make available to incident investigators on request (Ezell et al., 2025). This paper takes the next question: what happens when the party who must request the information cannot compel it, cannot interpret it without cooperation from the party whose conduct is in issue, and is subject to the system's account of events in the meantime?

The gap across these conversations is institutional rather than merely technical. Logging, reviewability, contestability, visibility, and incident analysis all presume that records can be made available to the institution that needs them.

## 3. Research design

The paper is a conceptual, doctrinal, and case-based study with a small documentary coding pilot. It does not test statistical hypotheses. It advances three propositions and examines them through sources that can separate legal authority, institutional access, evidentiary fragmentation, and agent-specific complications.

**Proposition 1 (P1):** computer evidence depends on institutional conditions as well as technical generation. A record becomes usable evidence only when an institution can preserve, access, test, interpret, and act on it.

**Proposition 2 (P2):** evidentiary fragmentation can arise without AI-specific properties. The relevant structure is not merely split custody, but a failure of timely authority and capacity to assemble, validate, interpret, and act on partial evidence.

**Proposition 3 (P3):** agent-mediated deployment intensifies evidentiary fragmentation because the governance object is the model-in-harness: model, execution environment, tools, permissions, verification loop, and organization.

Four methods examine these propositions. The doctrinal analysis establishes that control-conditioned duties exist in binding law and sit beside a common-law presumption that computer-generated evidence is reliable unless challenged. The Horizon case study tests P2 as a discriminating non-AI case: it shows what split

custody and limited access cost in a setting without the AI-specific properties usually blamed for evidentiary failure. The known-outcome GST test asks whether those conditions can be identified before deployment. The incident-record pilot asks what public AI failure records disclose about access to underlying evidence independently of the Horizon case.

The design is hybrid: doctrinal analysis, discriminating-case selection, known-outcome instrument testing, and documentary coding are established methods, and their combination is not presented as new. The combination is used because the research question requires legal analysis, a case that separates rival explanations, an operational instrument, and a small check on what public AI incident records make visible.

The propositions would be weakened by three findings. P1 would be weakened if retained computer records routinely became institutionally usable without meaningful access, validation, interpretation, or authority to act. P2 would be weakened if split custody were marginal in relevant failures because records were usually held by a single actor with ordinary disclosure duties and no practical access barriers. P3 would be weakened if agent-mediated action could be reconstructed from model-provider records alone, without evidence about the harness, tools, credentials, and deployment workflow.

The theoretical status of the paper is deliberately bounded. The framework is developed inductively from legal and documentary cases rather than derived from an existing theory. The structure it describes - complementary evidentiary fragments held by parties with divergent interests, asymmetric access, and costly compelled production - has clear affinities with information asymmetry, incomplete contracts, and principal-agent relations. Connecting evidentiary fragmentation to those literatures is a next theoretical step rather than a claim completed here.

## 4. From traces to evidence

AI deployments generate routine records such as inputs, outputs, timestamps, and errors. Agentic systems add trajectory-level records: tool calls, retrieved context, credential use, and monitoring alerts.

Runtime traces are valuable. Without them, many governance tasks are impossible. But traces do not govern by themselves.

The institutional question is how traces move through an evidence pipeline. The pipeline has seven stages.

**Table 1. Evidence Pipeline for Agent Runtime Traces**

| Stage | Governance question | Agent-governance example |
|---|---|---|
| Generation | What trace, action, or record is produced? | Prompt, tool call, API request, database query, approval request, error, output. |
| Custody | Who controls, preserves, versions, or can delete it? | Model provider, application developer, deployer, vendor, cloud service, public |

| Stage | Governance question | Agent-governance example |
|---|---|---|
| | | agency. |
| Access | Who can see, export, request, subpoena, or contest it? | User, operator, auditor, regulator, court, affected party, procurement office. |
| Validation | Who can verify that it is complete, accurate, and relevant? | Technical evaluator, internal audit team, incident investigator, external assessor. |
| Interpretation | Who can turn it into a finding or accountability claim? | Agency official, regulator, tribunal, compliance officer, investigator. |
| Action | Who can pause, remedy, sanction, update, or require change? | Deployer, regulator, procurement authority, court, model provider. |
| Learning | Does the incident improve institutional capacity? | Revised procurement rules, better logging, incident-derived evaluations, updated permissions. |

**Notes:** *Source: author's synthesis of the evidence-pipeline concept developed from the computer-evidence, reviewability, contestability, and agent-governance literatures cited in Sections 2 and 3. Analysis: each row maps a trace-to-evidence stage to the institutional question and an agent-governance example used in the paper.*

The evidence pipeline is institutional. Technical richness alone does not settle custody, access, validation, interpretation, action, or learning. A thin record can sometimes support a narrow remedy if the institutional pathway around it is strong.

Fragmentation can occur at several pipeline stages. Records may be technically distributed without being institutionally fragmented if a competent actor has timely authority to assemble and test them. Conversely, even centrally retained records can be evidentially fragmented if affected parties cannot access them or no independent institution can validate and interpret them. Fragmentation is therefore a property of the evidence pathway, not merely of storage architecture.

*Moffatt v. Air Canada*, 2024 BCCRT 149, shows the low-complexity version of the pipeline. A passenger used Air Canada's website after a bereavement. The chatbot gave inaccurate information about whether bereavement fare treatment could be requested after travel. The passenger relied on that representation, booked travel, and later sought the fare adjustment. Air Canada refused and argued that it should not be liable for the chatbot's statement. The tribunal rejected that position, treated the chatbot as part of Air Canada's website, and awarded damages, interest, and fees.

The case is useful because it is simple. An automated interface produced an output. A user relied on it. A firm tried to separate itself from the automated representation. A dispute forum compared the output with the firm's policy, assigned responsibility, and ordered a remedy.

The Air Canada record made responsibility and recourse visible. The public decision, however, did not reveal the chatbot's design, the underlying logs, or any post-incident learning loop. The tribunal could resolve the individual dispute. It did not produce a

full institutional account of how the failure emerged or whether similar failures had occurred.

The harder problem appears when an institution must reconstruct action. Existing legal and administrative systems can sometimes convert an automated output into responsibility and remedy. Agentic systems require reconstruction of action trajectories.

A wrong answer and a wrong action can both produce operational consequences. What differs is what must be reconstructed to establish responsibility. For an answer, the record required is the output and the reliance placed on it. For an action, it is the path from instruction through tools, credentials, and permissions to a change in external state, and that path is held by more parties than the output ever was.

For agent-mediated action, the relevant evidence includes the instruction, retrieved context, tools used, credential scope, records modified, and points where a human or institution could have intervened.

## 5. Method one: legal and institutional analysis of computer evidence

Modern institutions depend on records produced by the same computer systems whose conduct may later be questioned. Courts, agencies, firms, and regulated sectors routinely treat digital records as evidence of transactions, account states, access events, and institutional actions. AI agents intensify this older dependency.

In England and Wales, the current law of computer evidence rests on a common-law presumption that computers were operating correctly at the material time unless evidence is adduced to the contrary. Section 69 of the Police and Criminal Evidence Act 1984 once imposed a statutory requirement concerning computer operation before computer-produced statements could be admitted. That provision was repealed by the Youth Justice and Criminal Evidence Act 1999, with the change coming into force in April 2000. In January 2025, the UK Ministry of Justice launched a call for evidence on whether the current approach to software-generated evidence remains fit for the modern world, citing the Post Office Horizon scandal as the immediate context (Ministry of Justice, 2025a, 2025b).

The fragility of that presumption is not a new observation. Existing computer-evidence scholarship has argued that the presumption is poorly fitted to modern software systems (Ladkin et al., 2020; Marshall et al., 2021; Mason & Seng, 2021; Thimbleby, 2026). This paper does not claim novelty for that critique. It asks what happens when the question "was the computer operating correctly at the material time?" no longer has a determinate object because the relevant action path is distributed across a model-in-harness and multiple organizations.

The presumption is powerful because modern institutions cannot function if every computer record must be proved from first principles. It also depends on contestability. A party who challenges the record must be able to identify what to challenge and obtain enough evidence to make the challenge meaningful.

P3 applies directly to that assumption. A model provider may see the model, a harness provider may control the scaffold, a deployer may bear the public-facing duty, and an affected person may suffer the harm. No single party may be able to reconstruct the whole path without cooperation from actors whose interests diverge.

In those settings, the presumption of correct operation may become a presumption that affected parties cannot practically test.

The EU AI Act points toward the same evidentiary hinge. Articles 19, 21, and 26 require providers and deployers of high-risk AI systems to keep automatically generated logs to the extent those logs are under their control (Regulation (EU) 2024/1689). The wording is sensible as legal drafting because actors can retain only what they control. It also exposes the fragile point.

The phrase "under their control" is not a minor compliance detail. It is the hinge between technical trace generation and institutional accountability.

In agentic systems, different actors may control the model, scaffold, tool gateway, logs, and deployment workflow. A duty conditioned on control may therefore leave the central evidentiary question unresolved: when agent-mediated action fails, which actor controls the trace needed to reconstruct the action path, and what happens when the answer is several actors, none of them completely?

## 6. Method two: Horizon as a discriminating case

Horizon was the Post Office's branch accounting system. It recorded transactions and branch account states for sub-postmasters operating Post Office branches. The scandal matters for this paper because Horizon-generated shortfalls were treated as reliable institutional records while affected sub-postmasters had limited access to the technical evidence needed to challenge them. *Bates No. 6* is therefore not used here as a general technology-failure story. The Horizon case has been read for AI-governance lessons before, including as a case of accountability sinks and responsibility voids in digitised decision-making (Brennan et al., 2026). This paper uses it differently: not as a source of lessons but as a case that discriminates between two explanations of evidentiary failure. Furthermore, Horizon is a particularly discriminating case because it is a public judicial inquiry into whether computer-generated records could bear the legal and institutional weight placed on them.

Horizon did not lack technical traces. It had technical traces whose custody, completeness, interpretation, and disclosure were themselves part of the dispute. The case is methodologically useful because it separates two explanations that are otherwise hard to distinguish in contemporary AI governance. One explanation attributes evidentiary failure to AI-specific properties such as opacity, stochasticity, autonomy, or model updating. A second locates the failure in a different structure: difficult-to-reconstruct decision paths, split custody of relevant records, and divergent interests among the parties who hold or need those records. Horizon had the second structure without the first.

The inference is limited but important. Horizon establishes neither sufficiency nor frequency. It demonstrates that the evidentiary failure mechanism examined here can

arise without AI-specific properties. If evidentiary failure occurred in Horizon without opacity, stochasticity, autonomy, or model updating, then those features cannot be the whole explanation. The case is selected because it supports a mechanism and necessity claim, not a frequency claim.

The central public record for this section is *Bates & Others v Post Office Ltd (No. 6: Horizon Issues)* [2019] EWHC 3408 (QB), the judgment in which Fraser J determined the Horizon reliability issues after a full trial. *Hamilton & others v Post Office Limited* [2021] EWCA Crim 577 supplies the consequence layer: what happened when Horizon-derived evidence entered criminal proceedings. *Bates No. 6* shows what a full evidentiary inquiry required after the fact. The governance question is what would have been required to make that kind of inquiry possible without years of denial, group litigation, and criminal appeals.

The agreed Horizon Issues concerned four evidentiary questions: whether defects could create apparent discrepancies; what transaction data and reports were available to Post Office and sub-postmasters; whether remote access was possible; and whether branch data could be inserted, edited, deleted, or rebuilt without sub-postmaster knowledge or consent. These questions are set out in Horizon Issues 1, 3, 6, 13, and 14, reproduced at the beginning of *Bates No. 6* (para. 18). They make the case a judicial inquiry into whether computer-generated records could bear the institutional weight placed upon them.

**Technical records only create the possibility of evidence**

A record becomes usable evidence only when four conditions hold together: custody, completeness, accessibility, and interpretive authority. Someone must be able to obtain the record. The record must contain enough information to reconstruct the relevant state or action. The parties who need the record must have a pathway to reach it. An institution must be able to interpret the record without simply adopting the view of the party whose conduct is in dispute.

Horizon exposed failures across all four dimensions.

The first dimension was completeness. Horizon had internal technical records that should have made defects easier to investigate. Two categories mattered especially: PEAKs and KELs. PEAKs recorded incidents or problems. KELs, the Known Error Logs, identified known recurring issues. In *Bates No. 6*, the experts agreed that PEAKs and KELs were useful for understanding bugs, errors, and defects, while also recognizing that they were not comprehensive. KEL creation or updating had to be authorized before wider visibility. The records were real, useful, and incomplete at the same time (paras. 573-574).

The second dimension was custody. Claimants sought the Known Error Log in 2016. The Post Office resisted disclosure on relevance and proportionality grounds. Fraser J found the suggestion that the KEL was irrelevant wrong and without rational basis. The Post Office later pleaded that the KEL was maintained by Fujitsu and was not in Post Office's control. The court recorded that disclosure had been resisted on both control and content grounds, and Fraser J pushed back on the idea that a Fujitsu-maintained document was necessarily outside Post Office control (paras. 575-586).

The control point should be stated carefully. Post Office's control argument did not ultimately defeat disclosure. The sharper finding is that obtaining the relevant record required litigation, expert processes, judicial pressure, and time. The eventual reconstructability of a failure does not establish that its evidence was institutionally usable when consequences were imposed. Control-conditioned duties may be formally enforceable and practically inaccessible at the point when an affected person first needs the evidence.

The third dimension was accessibility. Audit data was treated in *Bates No. 6* as superior to management information for reconstructing what had occurred. Fraser J accepted that audit data did not need to be consulted for every transaction correction. In serious disputes between the Post Office and a sub-postmaster about branch accounts, however, he found that audit data should be consulted (paras. 909-911). The better evidence existed, but the operational routine did not require its ordinary use by the affected party.

Sub-postmasters could access some reports and logs, but only within limits. The judgment records that identifying causes of apparent discrepancies required cooperation between Post Office staff and sub-postmasters. Because the reports and data available to sub-postmasters were limited, their ability to investigate was correspondingly limited. The experts agreed that sub-postmasters could not themselves identify shortfalls, identify their causes, or properly identify the transactions recorded on Horizon (paras. 998-1000).

The fourth dimension was interpretive authority. *Bates No. 6* did more than collect technical records. It adjudicated competing interpretive postures. Post Office's expert reasoning often worked from the top down: treating Horizon as generally robust and asking whether a claimant had identified a specific defect that explained a discrepancy. The claimants' expert reasoning worked from the bottom up: examining defect records, audit data, remote access, and operational pathways through which discrepancies could arise. The court had to decide which records existed and how those records should be interpreted.

The competing interpretations made the proceeding an evidentiary inquiry rather than a data retrieval exercise.

**Logs without action-level reconstruction do not settle responsibility**

The Horizon judgment found that bugs, errors, and defects were not hypothetical. Fraser J concluded that bugs, errors, or defects could cause apparent discrepancies or shortfalls and undermine reliability, and that this had happened on numerous occasions. He also found a significant and material risk of branch accounts being affected by bugs, errors, and defects during relevant periods. The experts agreed that errors in data recorded within Horizon had occurred and led to financial discrepancies (paras. 967-970 and 980-982).

The defect findings matter because affected users were not merely raising speculative objections. Actual defect pathways existed. The problem was that defect histories did not become usable evidence soon enough for the people on whom the system's account states were imposed.

Fujitsu could alter branch accounts without sub-postmaster knowledge or consent. The relevant Horizon Issue asked whether Post Office or Fujitsu could insert, inject, edit, delete, implement fixes affecting, or rebuild branch transaction data, including without sub-postmaster knowledge or consent. For Fujitsu, Fraser J answered yes across the relevant permutations. Fujitsu witnesses accepted that injected transactions could appear as though the sub-postmaster had done them. Post Office also had a Global User role allowing transaction injection in branch, though physically present use made the posture different from Fujitsu remote access (paras. 1001-1006).

The evidentiary question is who or what changed the record, under whose authority, and whether the affected actor could see the difference.

Permission controls existed, but the logs could not reconstruct action. Fraser J found that roles were very wide and not properly controlled. Powerful Fujitsu roles were widely available and effectively unaudited. Privileged-user logs did not exist before 2009. From 2009 to July 2015, logs recorded log-on and log-off events but not user action. Even after July 2015, the experts agreed the logs were not useful evidence about remote access because of their content. The experts could not give confident evidence on frequency because of Fujitsu record-keeping (paras. 1007-1014). The weakness was also a separation-of-duties failure: actors who could alter branch records were inside the evidentiary system whose records would have to establish those alterations.

A log that records presence without action can establish who had access, but not what they did with it; responsibility turns on the second question.

The privileged-user finding carries directly into AI-agent governance. A deployment log must establish the action path at the level the accountability task requires: who authorized the task, which tool was called, which credential identity and scope were used, what state changed, and which approval gate was triggered or bypassed. Log existence is the beginning of the question, not the end.

**Contestability was missing at the point of institutional consequence**

Horizon also exposed a workflow problem. The judgment records that if a discrepancy appeared, a sub-postmaster could call the helpline or use dispute processes outside Horizon, but Horizon itself did not provide a way to mark a discrepancy as disputed in the system state. The accounting period also had to be rolled over. In practice, the disputed branch account state could continue to operate institutionally while the dispute was handled elsewhere (paras. 1017-1024).

The workflow finding is the paper's strongest contestability finding. Complaint was possible outside the system, but contestation did not alter the operative status of the record inside the workflow. The system-generated state could become the basis for debt, discipline, termination, civil claims, or prosecution before the affected party had meaningful access to the evidence needed to challenge it.

The retrospective Governance Stress Test therefore asks diagnostic questions with defined answer spaces. When an affected party disputes a system-generated state, what happens to that state pending resolution? Is it suspended? Annotated and still

operative? Operative and unmarked while dispute is handled outside the system? Or is there no mechanism? Horizon falls into the third category, with elements of the fourth: the dispute could be pursued outside the system, but the record's operative status was not transformed inside the workflow.

A second question follows. Does contestation pause institutional consequences created by the system state? The roll-over requirement meant that branch accounts had to be accepted before the next accounting period. That workflow converted a disputed state into an operative obligation before the evidentiary dispute had been resolved (paras. 1017-1024).

The consequence layer appears in *Hamilton*. The Court of Appeal addressed criminal convictions in which Horizon evidence, disclosure failures, and prosecutorial conduct interacted. For this paper, *Hamilton* is less important for the reliability facts than for the burden that unreliability imposed on defendants. The court held that failures of investigation and disclosure prevented appellants from challenging the reliability of Horizon data effectively, and later described Post Office as having treated an unreliable system-generated shortfall as an incontrovertible loss while defendants were denied disclosure capable of undermining the prosecution case (*Hamilton*, paras. 123 and 137).

The governance lesson is that a system can generate technical records and still fail as evidence infrastructure. Technical observability has not become institutional usability when an affected person cannot reach the record, the record cannot reconstruct action, the responsible organization can treat the system state as presumptively authoritative, or institutional learning arrives only after harm has compounded.

## 7. What AI agents change

Horizon shows that the evidentiary failure mechanism examined here can arise without AI-specific properties. AI agents still add important complications.

The AI-agent extension applies the Horizon mechanism before the paper turns to the stress-test instrument.

The first complication is trajectory. A conventional automated decision may produce an output from an input. An agent produces an action path. That path may include planning steps, tool calls, retrieved documents, API requests, credential use, and human approvals. The failure may sit in any part of that path.

The second complication is harness dependence. Zhang and co-authors argue that long-horizon agent comparisons are incomplete without disclosure of the execution harness: the layer that manages context construction, tool interaction, validation, retry logic, and stopping decisions (Zhang et al., 2026). Their methodological point reinforces GovAI's non-model capability-gains argument (Goemans et al., 2026). The same model can perform differently depending on the scaffold and execution environment around it. Agent capability is therefore not a property of the model alone. It is produced by the model-in-harness.

The governance object is not the model in isolation, but the model-in-harness: model, tools, state, credentials, verification, and organizational authority.

Harness dependence matters for both evaluation and deployment. If benchmark scores and capability claims depend on the harness, governance cannot stop at model-level evaluation. It must ask what task environment is being measured, what credentials are available, what feedback loop verifies action, and who can intervene when a model-mediated process becomes institutional action.

The third complication is split custody. Model providers, application builders, cloud providers, evaluators, and deployers may each hold part of the evidentiary trail. The affected person may have no access to any of the underlying records. This is the vendor-operator boundary in evidentiary form.

The fourth complication is task verifiability. Agents look stronger where tasks are tightly verifiable, such as code migration or test-suite improvement, and weaker where delegated work requires open-ended judgment. The CRUX shadow-evaluation work illustrates the distinction: in two early case studies, agents were given unpublished AI research questions and asked to produce publishable work; the original authors reviewed and rejected the outputs (Kirgis et al., 2026). The governance-relevant result is not simply that agents failed, but that the evidence needed to assess delegated work depends on the task, harness, evidence loop, and expert review protocol.

Different kinds of delegated action therefore require different evidentiary objects. A coding agent in a repository may be judged through tests, diffs, tool logs, and reviewer comments. A policy agent drafting a regulatory memo may require source provenance and expert review. A public-service agent modifying a benefits record may require authority, notice, recourse, and action-level reconstruction.

The OpenAI/Hugging Face incident is useful here less as a cyber narrative than as an evidentiary structure. Public reconstruction sits across separate disclosures issued by different organizations about connected portions of one event. OpenAI's post describes the evaluation-side setting and response. Hugging Face's disclosure and technical timeline describe the Hugging Face-side incident, forensic reconstruction, and remediation. Each document covers the portion its publishing organization placed in the public record. Neither is a general evidence return to a third-party investigator, and the public record does not itself show a rule requiring either document to aggregate the other's evidentiary layer (Hugging Face, 2026a, 2026b; OpenAI, 2026).

The evidentiary question is what an investigator, regulator, or affected institution could reconstruct from that disclosure structure. The two accounts together make the event more knowable than either account alone. The structure still illustrates split custody in an AI-agent setting: evaluation records, infrastructure records, and forensic accounts do not automatically sit in one place. Agent governance begins where benchmark evaluation ends. Once an agentic system can inspect, write, execute, and interact with its environment, the test environment becomes part of the system whose evidence may later be needed.

Agent systems may also create the opposite of the Horizon problem. Horizon's records were incomplete, inaccessible, or too thin to reconstruct action. Modern multi-agent systems may produce a data deluge: tool calls, intermediate states, traces, and environment events. The evidentiary problem then shifts from absence to selection. Institutions need trace structures and interpretive routines that can distinguish signal

from noise without letting the operator or vendor decide unilaterally which parts of the trajectory matter.

The deluge case is the same problem appearing at a different stage of the evidence pipeline. Where Horizon failed at custody and access, an over-instrumented agent deployment may fail at validation and interpretation. The records exist, but no one outside the operating institution can establish which fragments of a long trajectory are relevant. The party that chooses which events count as the relevant trajectory may acquire evidentiary power even when the raw records are formally available.

## 8. Method three: known-outcome testing of the Runtime Agent Deployment GST

The operational output of this paper is a proposed Runtime Agent Deployment Governance Stress Test (GST). A governance stress test asks whether the institution around a technology deployment can hold under realistic failure conditions. For AI agents, the relevant institution is the arrangement of authority, tools, credentials, logs, and incident routines that allows an agentic system to act in a real workflow.

A model audit may examine performance, documentation, security, or compliance. An impact assessment may identify expected harms. A benchmark may measure capability under a task setting. Incident analysis may reconstruct what happened after failure. The GST asks a different decision-facing question: what must the institution be able to demonstrate before it authorizes an AI agent to act?

The evidence pipeline describes how traces become evidence. The GST turns that pipeline into a deployment test. It asks whether the institution can specify the system and evidentiary object, inspect the action path, let affected parties or responsible staff contest a system-mediated state, govern intervention and responsibility across actors, and learn from failure.

The core question is:

> *What must an institution be able to demonstrate before it authorizes an AI agent to act rather than merely answer?*

The stress test has five areas. The areas below are diagnostic: they ask what an institution can establish. A stress test in the stronger sense places a deployment under a known-outcome or simulated incident and observes whether those capacities hold. The Horizon retrospective is a first version of that stronger test.

**Table 2. Runtime Agent Deployment Governance Stress Test: Five Diagnostic Areas**

| Area | Diagnostic question | Agent-deployment examples |
|---|---|---|
| Specify | Has the institution defined the system, task, authority, boundary, and evidentiary object? | What can the agent access? What can it change? Which model, harness, credentials, workflow, and version records are in scope? |
| Inspect | Can the institution observe and | Are tool calls, API requests, credential identity |

| Area | Diagnostic question | Agent-deployment examples |
|---|---|---|
| | reconstruct system behavior at the level needed for accountability? | and scope, approvals, errors, and state changes logged in integrity-verifiable, causally linked, searchable form? |
| Contest | Can affected parties or responsible staff challenge a system-mediated state before it becomes irreversible or highly consequential? | Can a record be marked disputed? Can underlying logs be obtained and filtered without relying only on the deployer's proprietary interpretation? Is there notice, review, and correction? |
| Govern | Can the institution intervene, constrain, suspend, contain, or assign responsibility across actors? | Which actions are irreversible? Are destructive or irreversible actions gated? Who owns incident response? |
| Learn | Do incidents, disputes, and evaluations improve institutional capacity? | Do failures produce revised procurement rules, updated evals, changed permissions, better logs, or staff learning? |

**Notes:** *Source: author's Runtime Agent Deployment Governance Stress Test, developed from the evidence pipeline in Table 1 and the Horizon known-outcome analysis in Sections 6 and 8. Analysis: each diagnostic area asks whether a deployment institution can demonstrate the corresponding capability before authorizing agent-mediated action.*

The Horizon retrospective test sharpens the instrument. If the GST would have passed Post Office Horizon in 2010, the instrument would be too weak for the failure mode this paper studies. Horizon would fail or become unanswerable across the five areas. The evidence boundary between Post Office, Fujitsu, audit data, known-error records, and remote access was not specified in a way affected parties could rely on. Audit data and error records existed, but the best evidence was not routinely consulted in disputes. Sub-postmasters could not meaningfully dispute the system-generated state inside the workflow before it became institutionally consequential. Authority, responsibility, and technical knowledge were split across Post Office and Fujitsu. Learning did not reliably change the default attribution posture or protect users soon enough.

The retrospective test also produced an instrument lesson. Gen 1 of the Governance Stress Test focused on authority, action safety, evidence, intervention, and reconstruction. That structure captured important features of agent deployment, but it did not separately anticipate contestability. Horizon exposed the miss. A system-generated state could be disputed outside the workflow while remaining operative inside it.

Gen 2 therefore uses five revised areas: specify, inspect, contest, govern, and learn. Gen 3 is the agent-specific extension: model and harness versioning, credential scope, runtime permissions, irreversible-action controls, and verification protocol. The paper does not claim that Gen 2 is validated because Horizon fails it. The methodological claim is specific: a known-outcome case can expose blind spots in a governance stress test, and a useful instrument must be able to fail when tested against a known failure.

For AI agents, the stress test should include specific runtime controls.

**Table 3. Agent-Specific Runtime Controls for the GST**

| Control area | Questions |
|---|---|
| Credential scope | Are credentials scoped, temporary, conditional, and revocable? Can the agentic system access production systems, and under what authorization? |
| Irreversible-action control | Are destructive or irreversible commands blocked by default? Are dry runs, command previews, approval gates, containment paths, and verified backups required? |
| Environment boundary | Are test, evaluation, staging, and production environments separated? Can the agentic system reach out-of-scope resources, including through external network paths? |
| Log integrity | Are tool calls, retrieved context, API calls, approvals, errors, and outputs logged with enough detail to reconstruct action? Are logs append-only or tamper-evident, causally ordered across distributed components, searchable, and protected from alteration by the actors being logged? |
| Reconstruction and versioning | Can an auditor distinguish model failure from harness failure, user error, credential misuse, or environment exposure? Are the model, harness, configuration, and tool versions identifiable? |

**Notes:** *Source: author's Gen 3 extension of the Runtime Agent Deployment GST, informed by the AI-agent extension in Section 7 and the Horizon known-outcome test in Section 8. Analysis: the table translates the GST into runtime-control checks for credential scope, irreversible action, environment boundaries, log integrity, and reconstruction/versioning.*

Approval gates are part of the institutional architecture of delegated action. They define which actions require human authorization, which actions must be blocked by default, and which records must exist before the institution treats an agent-mediated state as authoritative.

The GST has corresponding limits. Horizon can show that the evidentiary failure mechanism examined here can arise without AI-specific properties, but it cannot show how often fragmentation occurs in AI deployments. The incident-record pilot can show what selected public records contain, but not what the underlying incidents fully involved. The Runtime Agent Deployment GST is therefore a prototype instrument, not a validated standard.

## 9. Method four: what public incident records make knowable

The fourth method is a small incident-record coding pilot. The unit of analysis is the public incident record, not the underlying real-world incident. That distinction avoids representativeness claims the sample cannot support. The pilot used a dated bulk export from the AI Incident Database; later versions of the dataset may differ.

The first pass coded 30 AI Incident Database records for four fields: operating institution, technical-record custodian, affected-party access to the underlying record, and whether reliability of the record was contested (Butterfly Labs, 2026; McGregor, 2021).

The records were selected from the AI Incident Database/Butterfly Labs CSV export downloaded on 31 July 2026. The pilot excluded arXiv or research-only records and entries classified as low-severity in the source dataset, and preferred records with named developers, deployers, or harmed parties. The sample is therefore purposive and information-rich rather than representative. Codes distinguish affirmative indication from absence and silence: *Yes* means the field is affirmatively indicated; *No* means the public record affirmatively indicates absence; *Cannot determine* means relevant actors or disputes appear but the field cannot be resolved from public documentation; *Not addressed* means the public record does not discuss the field.

**Table 4. Public Incident-Record Coding Results**

| Field | Yes | No | Cannot determine | Not addressed |
|---|---|---|---|---|
| Operating institution identified | 30 | 0 | 0 | 0 |
| Technical-record custodian identifiable | 14 | 0 | 16 | 0 |
| Affected-party access indicated | 5 | 1 | 12 | 12 |
| Reliability of the record contested | 15 | 5 | 10 | 0 |

**Notes:** *Source: AI Incident Database/Butterfly Labs CSV export downloaded 31 July 2026. Analysis: n = 30 purposively selected public incident records coded for operating institution, technical-record custodian, affected-party access, and whether reliability was contested. Yes = affirmatively indicated; No = affirmatively absent; Cannot determine = relevant issue present but unresolved in public documentation; Not addressed = field not discussed.*

The pilot's contrast is simple: all 30 records identify an operating institution, while only five indicate whether an affected party could reach the underlying record. Those five are all court or tribunal records, including the Arizona probate appeal, CPS extradition filings, *Lexos Media v Overstock*, *Amarsingh v Frontier*, and the NCLT precedents matter. In this sample, affected-party access becomes visible in the public record only once the matter has entered litigation or adjudication.

The claim is about public evidence records, not the frequency of AI failures. When AI failure becomes publicly visible, the public record may still omit the information needed to reconstruct technical custody, access, validation, and contestability.

The pilot therefore offers a small empirical complement to the Horizon/KEL point without overstating it. Formal access may exist somewhere in the legal system, but public records often make that access visible only after the matter has become a court or tribunal problem. For governance, that is late. An evidence system that works only after litigation has begun may be formally enforceable and practically inaccessible at the point of institutional consequence.

## 10. Implications for AI governance

The paper's implications are focused rather than a general AI governance checklist.

First, logging requirements should be evaluated by reconstructability, not log existence. A retained record matters only if it can establish the action path at the level

needed for the decision: which system acted, under what authority, with which tools and credentials, what state changed, and which approval or exception route applied. For agentic systems, the harness, permissions, and verification loop are part of the governed object.

Second, evidence duties must account for control across organizational boundaries. Horizon shows that a formally enforceable duty may still be practically inaccessible when the relevant record sits across vendor, operator, infrastructure, or expert-control relationships. The question is not simply whether a rule requires production, but whether the institution that needs the evidence can compel, interpret, and use it at the moment of consequence.

Third, contestability must address the institutional status of a disputed machine-generated state while the dispute remains unresolved. A challenge process outside the workflow is not enough if the system-generated state continues to structure debt, discipline, service access, or legal exposure inside the workflow. The governance problem becomes acute when institutional consequences move faster than the evidence pathway required to contest them. In-workflow contestability requires a way to mark, suspend, annotate, review, or otherwise condition the institutional force of the record while its evidentiary basis is unresolved.

These findings have practical consequences for procurement, auditing, and oversight. Public agencies and regulated organizations can require integrity-verifiable action logs, evidence-preservation duties, versioning, approval controls, and responsibility allocation across vendors and operators before deployment. Cooperative access remains valuable, but it should not be confused with inspection authority; a public institution that cannot compel or vary the relevant evidence remains dependent on private cooperation.

Human oversight also has to be operational rather than symbolic. A human at the final decision point does not solve an upstream evidence problem if the underlying trajectory cannot be reconstructed or contested. Oversight requires preserved records, intelligible action paths, authority to intervene, and routines that convert failure into institutional learning.

The larger research agenda is the institutional governance of delegated AI action: what institutions must be able to reconstruct, contest, and authorize as AI systems move from producing outputs to exercising institutional force. This paper develops the reconstruction problem within that broader agenda.

The immediate next step is more specific. The Runtime Agent Deployment GST should be applied to a small set of cases: Horizon as the non-AI discriminating case, OpenAI/Hugging Face as the evaluation-environment case, Air Canada as the automated-output responsibility case, CRUX as the verifiability case, and one public-sector deployment case where affected-party access and recourse are central.

Together, these cases would test whether the evidence-pipeline framework travels across variation in automation type, task verifiability, organizational boundaries, and institutional consequence.

The final claim is deliberately constrained. AI agents create new governance problems, including stochasticity, opacity, delegated action, and dynamic environments. The institutional failure mode begins earlier, when a system exercising institutional force produces records that institutions cannot use.

Agent governance therefore turns on whether institutions can convert traces into usable evidence before agent-mediated action acquires institutional force.

## References


Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2023). Contestable AI by design: Towards a framework. *Minds and Machines, 33*, 613-639. https://doi.org/10.1007/s11023-022-09611-z

Almada, M. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems. Proceedings of ICAIL 2019. https://doi.org/10.1145/3322640.3326699

*Bates & Others v Post Office Ltd (No. 6: Horizon Issues)* [2019] EWHC 3408 (QB). Searchable Inquiry copy: https://postofficeinquiry.dracos.co.uk/projects/evidence/RLIT0000005/

Brennan, N. M., Edgar, V. C., & Gorman, L. (2026). Accountability sinks, accountability gaps and technological injustice: The case of the British Post Office/Horizon IT scandal. Accounting, Auditing & Accountability Journal. Advance online publication. https://doi.org/10.1108/AAAJ-11-2025-8532

Brundage, M., et al. (2026, January 16). Frontier *AI auditing: Toward rigorous third-party assessment of safety and security practices at leading AI companies*. GovAI. https://www.governance.ai/research-paper/frontier-ai-auditing-toward-rigorous-third-party-assessment-of-safety-and-security-practices-at-leading-ai-companies

Butterfly Labs. (2026). AI Incident Database CSV export used for pilot coding [Data set]. Downloaded July 31, 2026. https://incidents.butterflylabs.org/api/export/incidents.csv

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024, June 5). *Visibility into AI agents*. GovAI. https://www.governance.ai/research-paper/visibility-into-ai-agents

Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025, January 17). *Infrastructure for AI agents*. GovAI. https://www.governance.ai/research-paper/infrastructure-for-ai-agents

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable automated decision-making: A framework for accountable algorithmic systems. *Proceedings of FAccT 2021*. https://arxiv.org/abs/2102.04201

Cobbe, J., Veale, M., & Singh, J. (2023). Understanding accountability in algorithmic supply chains. *Proceedings of FAccT 2023*. https://arxiv.org/abs/2304.14749

Ezell, C., Roberts-Gaal, X., & Chan, A. (2025, August 19). *Incident analysis for AI agents*. GovAI. https://www.governance.ai/research-paper/incident-analysis-for-ai-agents

Goemans, A., Altman, D., Dreksler, N., Freund, J., Gandhi, M., Wang, Z., Cogan, S., Krier, S., Brady, D., Ho, L., & Dafoe, A. (2026, May 1). *Comprehensive AI governance requires addressing non-model capability gains*. GovAI. https://www.governance.ai/research-paper/comprehensive-ai-governance-requires-addressing-non-model-capability-gains

*Hamilton & others v Post Office Limited* [2021] EWCA Crim 577. https://www.judiciary.uk/judgments/hamilton-others-v-post-office-limited/

Henin, C., & Le Métayer, D. (2022). Beyond explainability: Justifiability and contestability of algorithmic decision systems. *AI & Society, 37*, 1397-1410. https://doi.org/10.1007/s00146-021-01251-8

Hugging Face. (2026a, July). *Security incident disclosure - July 2026*. https://huggingface.co/blog/security-incident-july-2026

Hugging Face. (2026b, July 27). *Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident*. https://huggingface.co/blog/agent-intrusion-technical-timeline

Kirgis, P., et al. (2026, July 30). *Can AI agents conduct open-ended AI research? Early evidence from two case studies*. arXiv. https://arxiv.org/pdf/2607.27191

Ladkin, P. B., Littlewood, B., Thimbleby, H., & Thomas, M. (2020). The Law Commission presumption concerning the dependability of computer evidence. *Digital Evidence and Electronic Signature Law Review, 17*, 1-14. https://doi.org/10.14296/deeslr.v17i0.5143

Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising contestability. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1-25. https://doi.org/10.1145/3449180

Marshall, P., Christie, J., Ladkin, P. B., Littlewood, B., Mason, S., Newby, M., Rogers, J., & Thimbleby, H. (2021). Recommendations for the probity of computer evidence. *Digital Evidence and Electronic Signature Law Review, 18*. https://doi.org/10.14296/deeslr.v18i0.5240

Mason, S., & Seng, D. (Eds.). (2021). *Electronic evidence and electronic signatures* (5th ed.). University of London Press. https://read.uolpress.co.uk/projects/electronic-evidence-and-electronic-signatures

McGregor, S. (2021). Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database. *Proceedings of IAAI-21*. https://incidentdatabase.ai/research/snapshots/

Ministry of Justice. (2025a, January 21). *Use of evidence generated by software in criminal proceedings: Call for evidence*. https://www.gov.uk/government/calls-for-evidence/use-of-evidence-generated-by-software-in-criminal-proceedings

Ministry of Justice. (2025b, January 21). *Use of computer evidence in court to be interrogated*. https://www.gov.uk/government/news/use-of-computer-evidence-in-court-to-be-interrogated

*Moffatt v. Air Canada*, 2024 BCCRT 149. https://decisions.civilresolutionbc.ca/crt/crtd/en/item/525448/index.do

OpenAI. (2026, July 21). *OpenAI and Hugging Face partner to address security incident during model evaluation*. https://openai.com/index/hugging-face-model-evaluation-security-incident/

Police and Criminal Evidence Act 1984, c. 60, s. 69 as enacted. https://www.legislation.gov.uk/ukpga/1984/60/section/69/enacted

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Thimbleby, H. (2026, January 9). *Law reform driven by understanding computers*. The Bar Council. https://www.barcouncil.org.uk/resource/law-reform-driven-by-understanding-computers.html

Vaccaro, K., Karahalios, K., Mulligan, D. K., Kluttz, D., & Hirsch, T. (2019). Contestability in algorithmic systems. *CSCW 2019 Companion*, 523-527. https://doi.org/10.1145/3311957.3359435

Youth Justice and Criminal Evidence Act 1999, c. 23, s. 60. https://www.legislation.gov.uk/ukpga/1999/23/section/60

Zhang, Y., Wang, J., Ge, Y., Xu, W., Hamm, J., & Reddy, C. K. (2026, May 7). *Stop comparing LLM agents without disclosing the harness*. arXiv. https://arxiv.org/abs/2605.23950