Operative-State Contestability

Agent-Mediated Action and the Institutional Conversion of Computer Evidence

Venkat Nadella

Working paper - 7 September 2026

Abstract

Contemporary agent systems increasingly produce technically reconstructable records of model calls, tool use, credentials, infrastructure events, and external actions. Recent work on observability, provenance, forensic readiness, runtime witnessing, proof-carrying actions, per-decision evidence, and preserved evidence bundles makes that record layer richer. This paper asks what remains after that progress is conceded. It identifies **operative-state contestability**: an institutional property whereby a relevant challenge can trigger a process capable of altering the status or consequences of an operative state while the evidentiary basis for that state is assembled and reviewed. Methodologically, it uses Horizon as a discriminating non-AI case, US public-benefits litigation as a comparative administrative line, the July-August 2026 OpenAI/Hugging Face/METR record as an agent-side contrast, and a capability-duty matrix and Runtime Agent Deployment Governance Stress Test prototype as diagnostic translations.

The Post Office Horizon litigation isolates the mechanism in a conventional non-AI accounting system: later reconstruction can coexist with weak contestability at the point where a machine-mediated account state becomes consequential. US public-benefits litigation shows a related structure in rule-based and data-matched administration. The July-August 2026 OpenAI/Hugging Face/METR record supplies the agent-side contrast: complex agent activity can be reconstructed across custodians, but authority over any disputed state remains an open institutional question. Technical reconstruction asks whether a later investigator, court, or expert can determine what happened. Institutional conversion asks whether that record can support contestation and status control while the disputed state remains governable, or remedy after consequences harden.

1 Introduction

The August 2026 public record of the OpenAI/Hugging Face incident is a useful place to begin because it shows that the record layer is not empty. OpenAI published a postmortem. METR and Redwood Research published an independent investigation based on mediated access to OpenAI-side records. Hugging Face published a technical timeline reconstructing roughly 17,600 attacker actions from sandbox, launchpad, and platform logs (Hugging Face 2026). Between them, the accounts make a complex agent incident partially legible across organisational positions.

The remaining question, therefore, cannot be framed as a simple absence of logs. The strongest contemporary agent-evidence work is already building forensic readiness, runtime witnessing, execution provenance, graded attribution, proof-carrying actions, and preserved evidence bundles. This paper investigates what follows when present systems can produce technically useful traces.

Horizon supplies the contrasting case: a conventional branch accounting system whose records became authoritative in contractual, civil, disciplinary, and criminal settings. It

was not an AI system. Years later, litigation and public inquiry made the record more visible. But affected sub-postmasters lacked timely evidentiary power over the shortfall states while those states structured repayment demands, suspension, termination, recovery, and prosecution. The institutional problem was whether computer records could become usable evidence for the actor and decision that mattered before the disputed state had already done institutional work. This paper examines that failure through the concept of operative-state contestability: the institutional capacity to alter a disputed state's status or consequences while its evidentiary basis is assembled and reviewed.

This paper therefore separates two processes that are often conflated. **Evidence formation** concerns whether a runtime record is generated, preserved, assembled, authenticated, and interpreted. **Accountability conversion** concerns whether that record can be used by a relevant actor, through an appropriate institutional pathway, to alter or remedy the consequences of a disputed state. A technically reconstructable record may still fail at conversion.

Existing work already covers the reliability of computer-generated records, distributed accountability, agent provenance, observability, forensic readiness, model-in-harness analysis (Zhang et al. 2026), and governance stress testing. This paper's claim is that a residual governance problem remains even where preservation form, cross-custodian verification, and interpretive scoping are increasingly addressed: reconstruction does not imply contestation, and contestation does not imply status control. Using the construct of operative-state contestability, the paper examines the gap through a discriminating non-AI case, a comparative administrative-law line, a contemporary agent-side contrast, a capability-duty matrix, and a prototype deployment diagnostic.

2 Related Work and the Residual Institutional Problem

Computer-evidence scholarship has long examined software-generated records, computer dependability, disclosure obligations, and the risk of treating complex systems as neutral witnesses (Mason and Seng 2021; Ladkin et al. 2020; Marshall et al. 2021). The Ministry of Justice's 2025 call for evidence on software-generated evidence in criminal proceedings confirms that the legal question remains live.

Contestability scholarship supplies a second baseline, examining contestability as design, justification, procedure, reviewability, and human intervention (Alfrink et al. 2023; Almada 2019; Lyons et al. 2021; Henin and Le Métayer 2022; Vaccaro et al. 2019; Cobbe et al. 2021, 2023). Those accounts are necessary, but they usually do not ask the status-control question posed in this paper: what happens to the operative state while the challenge is pending? Operative-state contestability is needed because ordinary contestability can describe a route to complain, a right to explanation, or human review without specifying whether the challenged state may keep producing consequence while evidence is assembled.

Agent provenance and forensics literature supplies the most important technical baseline. Chan and co-authors' visibility agenda treats agent identifiers, real-time monitoring, and activity logs as governance measures for agent risk, while later agent-infrastructure work asks what external protocols and systems are needed to shape agent interactions with institutions and other actors (Chan et al. 2024; Chan et al. 2025). This paper sits downstream of that agenda: visibility and infrastructure make evidence formation more plausible, but they do not by themselves determine who can use the evidence to alter a disputed operative state. *From Agent Traces to Trust* (Wang, Y., et al. 2026) treats execution provenance as a typed graph of agent execution. W3C PROV and PROV-DM supply the general provenance vocabulary. MCPRecon (Satter et al. 2026) reconstructs some Model Context Protocol activity from volatile memory. HANSARD (Sardianos et al. 2026), AUDITA (Du and Chen 2026), PCAA (Wang, Z. 2026), AIREP (Abak 2026), preserved evidence-bundle work (Alshammari and Bahsi 2026), and Janssen's runtime-records-to-legal-findings criterion then push from traces toward readiness, attribution, authority semantics, signed per-decision evidence, policy-relative sufficiency, and legally operative answerability.

The research question pursued in this paper is institutional. The paper does not claim priority over the record layer. HANSARD asks whether a future investigator can trust the record enough for forensic attribution. That is a hard and important question, especially where a multi-agent system may help produce the record later used to judge it. HANSARD's answer is architectural: specify readiness before operation, witness events at runtime, preserve commitments, align records to provenance structures, and grade the strength of attribution after the event. Such an architecture is designed to answer much of the Inspect problem in this paper.

AUDITA supplies a certified record-and-attribution layer with legal-facing ambition. PCAA specifies authority and approval semantics for certificate-bearing actions within heterogeneous runtimes. AIREP creates a per-decision record form whose governance verbs include release, block, defer, redact, escalate, and halt. Those verbs are close enough to the status language of this paper that the boundary must be explicit: AIREP's verbs are the runtime's own disposition of its own output, applied at or before release on the system's initiative. Operative-state contestability concerns the institution's response to a relevant actor's challenge after a state has become operative and started producing consequences. A runtime that defers its own output has not thereby given an affected actor power to contest a benefits reduction, debt, sanction, or account state already treated as authoritative.

One convergence across the reviewed record-layer literature is especially important. HANSARD, AUDITA, AIREP, and preserved evidence-bundle work converge on a design pattern: publish or share the verifiable commitment, gate the underlying content, and leave the mechanism for obtaining content to legal or institutional process. That convergence is the paper's central premise. Technical architectures increasingly address preservation form, cross-custodian verification, and interpretive scoping. The residual question is whether any relevant actor can alter the interim status of a disputed operative state, and who has remedial authority if the record later supports the challenge.

The novelty is therefore not interim status alone. Suspensive effect, stays, interim relief, and provisional measures are old institutional tools. Nor is the novelty that automated decisions should be contestable. The contribution, in the literature reviewed for this draft, is the coupling: the institutional status of a disputed machine-mediated state may depend on an evidentiary process that itself requires preservation, cross-system assembly, authentication, and interpretation before the relevant actor can use it. Reconstruction does not imply contestation, and contestation does not imply status control.

Similarly, legal and policy scholarship drawing lessons from the Horizon case is adjacent and cumulative. Lallie treats Horizon as a socio-technical failure and maps it against cybersecurity and digital-forensics frameworks, including evidence integrity, privileged access, and institutional independence (Lallie 2026). Brennan, Edgar, and Gorman analyse witness statements through accountability sinks and gaps, showing how digitised accounting weakened sub-postmasters' ability to demand and give accounts (Brennan et al. 2026). The distinctive use of Horizon in this paper is methodological and temporal: Horizon is used not as a first case of digital-evidence governance, but as a discriminating case showing that operative-state contestability can fail without AI-specific features such as stochasticity, model opacity, autonomy, or prompt sensitivity.

3 Research Design and Propositions

The study's research design is deliberately comparative and mechanism-oriented. It does not seek to estimate prevalence or claim that agent systems uniquely create the problem. It investigates three questions: whether a specific institutional mechanism can be isolated, whether that mechanism appears in adjacent administrative settings, and whether contemporary agent systems create conditions under which the mechanism may become more difficult to govern.

The propositions are correspondingly bounded.

P1. Technical reconstructability does not guarantee institutional conversion. A record may later become reconstructable enough for expert or judicial findings while remaining unavailable, uninterpretable, or institutionally ineffective during the consequence window.

P2. Operative-state contestability can fail without AI-specific technical properties. Horizon is selected because it isolates the mechanism in a conventional non-AI software system. If the failure appears there, opacity, stochasticity, autonomy, and model updating cannot be the full explanation.

P3. Agent-mediated deployments may intensify the problem where action trajectories cross organisational or technical boundaries, relevant records are distributed across custodians or generated through components or channels the acting system can influence, and no timely pathway joins access, interpretation, and authority over the disputed state. P3 is a specified hypothesis awaiting prospective testing. The OpenAI/Hugging Face record supports its technical premises but does not demonstrate affected-party institutional consequence.

The materials play different evidentiary roles in the paper's analytical approach: Horizon is the discriminating case; US public-benefits litigation supplies a comparative administrative line; OpenAI/Hugging Face supplies the contemporary agent-side contrast. The capability-duty matrix tests whether the residual pathway is already supplied by technical or legal instruments. The Runtime Agent Deployment Governance Stress Test (GST) translates the concept into a diagnostic for prospective deployment work.

4 Legal and Evidence Baseline

The paper uses "evidence" more broadly than courtroom admissibility, but the legal distinctions matter. English criminal law, English civil disclosure, administrative review, public-sector procurement, and US litigation do not ask the same evidentiary questions. A record can be admissible but weak in weight, discoverable but late, technically authentic but institutionally unusable, or sufficient for an operator's internal investigation but insufficient for an affected person's challenge.

The criminal line matters because computer-generated evidence has carried legal consequences. The Police and Criminal Evidence Act 1984 (PACE) s. 69 formerly imposed a statutory reliability condition for computer evidence in criminal proceedings; its repeal through the Youth Justice and Criminal Evidence Act 1999 s. 60 left courts to rely on ordinary evidentiary principles and the rebuttable common-law presumption that a computer producing evidence was operating correctly at the material time. The Ministry of Justice call for evidence on software-generated evidence in criminal proceedings shows that this settlement remains contested. Wexler's work on trade secrets in criminal justice supplies the US analogue: proprietary systems can structure criminal evidence while defendants struggle to inspect the technical basis of the record.

The civil line is different. *Bates No. 6* was a civil judgment in group litigation, not a criminal admissibility decision. Its value for this paper is not that civil disclosure rules directly governed every later use of Horizon evidence. Its value is that a court, after extensive adversarial process, could reconstruct enough about Horizon to show that earlier institutional reliance had outrun affected-party evidentiary power. That is a mechanism claim, not a transposition of civil disclosure doctrine into criminal law.

Civil procedure also shows why preservation law is powerful and late. CPR Part 31 and PD31B continue to matter outside the Business and Property Courts, while PD57AD governs disclosure in the Business and Property Courts subject to exclusions and made the disclosure pilot permanent from 1 October 2022. PD57AD expressly extends documents to server, backup, deleted, metadata, and embedded information, requires reasonable preservation of relevant documents in a party's control, and requires suspension of relevant deletion or destruction processes during proceedings (PD57AD paras. 2.5-2.6, 3.1, 4.1-4.2). That performs a preservation function analogous to litigation-hold practice. It

is also triggered by litigation or contemplated proceedings, which may arrive after the machine-mediated state has produced the very consequence later disputed.

This is the procedural baseline for operative-state contestability. Evidence law can reconstruct, compel, exclude, weigh, and remedy. Administrative law can invalidate rules or require lawful process. Procurement can impose record duties. Regulators can inspect within statutory powers. The gap appears when these mechanisms do not join early enough, across the right custodians, with authority to alter the disputed state while it remains consequential.

5 Core Definitions

5.1 Operative States

This paper concerns machine-mediated states that acquire institutional force.

An **informational output** is advice, text, explanation, prediction, or analysis that an institution may consider but does not itself treat as authoritative. A **provisional system state** is a flag, score, recommendation, draft status, anomaly marker, or proposed classification that may inform action but remains subject to ordinary review before consequence attaches. An **operative state** is a machine-mediated representation, classification, account condition, decision state, or external-system change that an institution treats as sufficiently authoritative to produce consequences for an actor or resource.

Examples include a branch account shortfall treated as a debt, a benefits calculation treated as reducing entitlement, a fraud flag treated as authorising recovery or sanction, a platform enforcement state treated as disabling an account, or an agent-initiated external-system change treated as an authorised act. The term is status-sensitive: an output becomes operative when institutional practice gives it consequence.

5.2 Operative-State Contestability

Operative-state contestability is an institutional property whereby a relevant challenge can trigger a process capable of altering the status or consequences of an operative state while the evidentiary basis for that state is assembled and reviewed.

The definition is narrower than contestability in general. Design properties, justification requirements, procedural affordances, lifecycle governance, and human intervention remain necessary. They do not by themselves answer this paper's question: what happens to the institutional status of the operative state while the challenge runs?

A system may allow a complaint, an appeal, a helpdesk ticket, a human review, or later litigation while the disputed state continues to operate as authoritative. In that situation the existence of a route to challenge does not establish operative-state contestability. The relevant question is whether the challenge changes what the state may do before the evidentiary question is resolved.

The portable test is:

> When a relevant actor raises a challenge, does the institution alter or formally specify the disputed state's permissible uses, consequences, or review status before the evidentiary question resolves?

The answer space should be specified before deployment:

Status during challenge	Meaning
Suspended or limited	The state cannot produce some or all consequences while review is pending.
Annotated but operative	The state remains consequential, but institutional users can see that it is disputed.
Operative and unmarked	The state remains consequential and appears undisputed in the workflow.

Status during challenge	Meaning
No challenge mechanism	There is no route by which the relevant actor can trigger status review.

These categories are not a hierarchy in which suspension is always required. Operative-state contestability is risk-sensitive. The appropriate interim status depends on consequence, reversibility, urgency, and risk of abuse. A practical rule of thumb is that suspension or hard limitation becomes more important as consequences become severe, irreversible, or difficult to compensate, while annotation or escalation may be sufficient where consequences are low, reversible, or already subject to rapid human review. Fraud-prevention systems, cybersecurity controls, critical-infrastructure safeguards, account-debt processes, disciplinary systems, and public-benefits determinations may require different interim postures. The governance requirement is that the posture be specified and institutionally available, not that every challenge automatically suspends every state.

The concept is easiest to keep distinct through a three-part sequence. **Reconstruction** asks whether someone can later determine what happened. **Contestation** asks whether a relevant actor can challenge the record, interpretation, or state. **Status control** asks whether that challenge changes what the disputed state may do while review is pending. The paper's central claim sits in the gap between the three: reconstruction does not imply contestation, and contestation does not imply status control.

Failure occurs when an operative state acquires institutional force before the relevant party can trigger a process capable of preserving and obtaining a sufficient record, testing its interpretation without exclusive dependence on the actor asserting the state, and altering or suspending the state's consequences where warranted.

The failure is temporal because consequence can accumulate between the moment a state becomes operative and the moment evidence becomes usable. It is institutional because the question is not only whether a log exists, but who can trigger preservation, obtain the record, interpret it without exclusive reliance on the party asserting the state, and alter the state's operative status in time.

5.3 Consequence Window

The **consequence window** is the interval between the moment a machine-mediated state becomes operative and the point at which its consequences have become practically difficult to prevent, reverse, or remedy. It is not fixed by clock time alone: a branch account roll-over, benefits reduction, fraud recovery, account suspension, disciplinary action, prosecution decision, or irreversible external-system change may each harden on a different institutional timeline.

The consequence window is the period in which operative-state contestability matters most. After that window, evidence may still support correction, compensation, appeal, sanction, or institutional learning. But delayed remedy is not the same as status control. The question is whether the challenged state can be limited, marked, escalated, or suspended while it is still doing institutional work.

The analytic pressure rises as evidentiary latency approaches or exceeds the consequence window. Where a complete-enough record becomes usable before consequences harden, ordinary review may suffice. Where the record becomes usable only after the state has structured debt, deprivation, sanction, exclusion, or legal exposure, retrospective remedy substitutes for contestability rather than satisfying it.

Structural fragmentation occurs when complementary records lie across custodians, systems, trust domains, vendors, deployers, or legal actors. A model provider may hold model-side logs, a platform may hold account or infrastructure logs, a deployer may hold workflow records, and an affected party may hold only the output or consequence.

Temporal fragmentation occurs when the record can eventually be reconstructed, but not before institutional consequences attach or harden. This is the distinctive form on which

the paper focuses. A record that arrives after debt has been enforced, benefits reduced, an account disabled, a prosecution advanced, or an irreversible action completed may be technically useful and institutionally late at the same time.

Fragmentation is therefore not the same thing as distributed storage. Distributed records are not necessarily fragmented if a competent actor can assemble, test, and act on them within the consequence window. Centrally retained records can still be fragmented if no actor with relevant authority can obtain and use them in time.

A runtime record becomes institutionally usable evidence when an authorised actor can preserve, obtain, authenticate, interpret, and use it for the accountability task at issue. This paper uses evidence more broadly than admissibility in court; it includes use by a regulator, auditor, agency, tribunal, ombudsman, investigator, procurement authority, deployer, affected party, or court.

The relevant standard is not exhaustive reconstruction. It is **task-relative institutional sufficiency**: enough evidence for which institutional task, decided by which actor, with what consequence, within what time, and with what authority over the disputed state. This differs from policy-relative evidence sufficiency in preserved evidence-bundle work, where the question is whether a verifier can accept a bundle under a selected policy profile.

The same record can be sufficient for one purpose and insufficient for another. A trace may let an operator debug a tool call without letting an affected party contest a debt. A signed event bundle may let a verifier confirm a policy-required message without letting a tribunal decide whether a benefits reduction should remain operative during statutory challenge.

Institutional usability does not require direct affected-party access to raw telemetry in every case. Privacy, security, trade-secret, and abuse concerns may make direct access inappropriate. The relevant access question is whether the contestation process can obtain and test the evidentiary basis without depending exclusively on the actor asserting the operative state.

The paper distinguishes:

Access type	Role in the evidentiary pathway
Public access	Records or explanations are publicly released, but may be partial, curated, or non-compellable.
Mediated independent access	External investigators receive structured access under conditions set by one or more custodians.
Regulatory access	A public authority can compel or inspect records under statutory powers.
Judicial or compelled access	A party can obtain records through litigation, disclosure, subpoena, or court order.
Representative access	An auditor, ombudsman, advocate, union, class representative, or technical expert accesses records on behalf of affected actors.

The retained Supplement A operationalises this point by disaggregating access by object, holder or audience, mechanism, timing, and usability. That prevents a public output, an operator dashboard, a court-supervised technical record, a regulator's compulsory access power, and a verifiable commitment from being treated as the same institutional condition.

Interpretive independence should be understood modestly. It does not mean that every institution can interpret complex technical records without assistance from operators or vendors. It means the institution is not forced to adopt the disputed operator's account as the exclusive interpretation of the record. The threshold is interpretive independence sufficient for the institutional task.

6 Horizon as the 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. It is not used here because it was the first or only digital evidence scandal. It is used because it isolates the paper's mechanism in a conventional non-AI transaction system. If a failure of operative-state contestability can arise without stochastic outputs, model opacity, autonomous planning, model updating, or prompt-sensitive behaviour, then those AI-specific features cannot be the whole explanation for the evidentiary problem.

The public record is unusually strong. *Bates & Others v Post Office Ltd (No. 6: Horizon Issues)* is a civil judgment after a full trial on Horizon reliability. *Hamilton & others v Post Office Limited* supplies the criminal consequence layer, showing what happened when Horizon-derived shortfalls supported prosecutions and convictions. The Post Office Horizon IT Inquiry supplies witness evidence and institutional history. Recent scholarship now also treats Horizon as an accountability and evidence-governance case: Lallie maps Horizon against cybersecurity and digital-forensic frameworks, and Brennan, Edgar, and Gorman code witness evidence to show how digitised accounting weakened the ability of sub-postmasters to give and demand accounts (Lallie 2026; Brennan et al. 2026). This paper's narrower use is different. Horizon is a discriminating case for operative-state contestability.

The first point is that Horizon did not simply lack records. It had internal technical records whose custody, completeness, accessibility, and interpretation were themselves contested. PEAKs recorded incidents or problems. KELs recorded known errors. Audit data could be more probative than ordinary management information. Remote-access and privileged-user records could have mattered to attribution (*Bates No. 6*, paras. 573-586, 909-911, 1001-1014). The difficulty was not that nothing existed. It was that the record capable of challenging an operative branch-account state was incomplete, late, unevenly accessible, and dependent on institutional actors whose own position was in dispute.

Bates No. 6 makes that structure visible. The agreed Horizon Issues asked 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 Post Office or Fujitsu could insert, edit, delete, rebuild, or otherwise alter branch transaction data without sub-postmaster knowledge or consent (*Bates No. 6*, para. 18). Those questions are evidentiary questions before they are software questions. They ask whether a machine-mediated state could bear the institutional weight being placed on it.

The record eventually became reconstructable enough for judicial findings. That eventual reconstruction is precisely why Horizon matters. Fraser J found that bugs, errors, and defects could and did cause apparent discrepancies or shortfalls, and that there was a material risk of branch accounts being affected by such defects (*Bates No. 6*, paras. 967-970, 980-982). He found that Fujitsu could alter branch transaction data without sub-postmaster knowledge or consent, and that some injected transactions could appear as if a sub-postmaster had performed them (*Bates No. 6*, paras. 1001-1006). He also found that privileged-user records were inadequate for reconstructing action: before 2009 they did not exist, from 2009 to mid-2015 they recorded log-on and log-off rather than user actions, and even later they were not useful evidence about remote access (*Bates No. 6*, paras. 1007-1014). A log existed in some periods, but it did not answer the accountability question that mattered.

Access was similarly partial. Sub-postmasters could see some reports and could raise issues through support channels, but *Bates No. 6* records that they could not themselves identify shortfalls, identify their causes, or properly identify the transactions recorded on Horizon (*Bates No. 6*, paras. 998-1000). Brennan, Edgar, and Gorman's witness-based study corroborates the access problem from a different dataset: sub-postmaster witnesses repeatedly described being unable to access the system and treating the helpline as a dead end (Brennan et al. 2026). Their study also records evidence that transaction corrections could have to be accepted before a branch could continue operating. That is not merely

poor access to information. It is an operative-state problem: the disputed accounting state continued to structure what the affected person had to do.

The strongest Horizon finding for this paper is therefore not "the software was unreliable." It is that the disputed state could remain operative while the evidentiary challenge was handled elsewhere. *Bates No. 6* records that a sub-postmaster could call the helpline or use external dispute processes, but Horizon did not provide a way to mark the discrepancy as disputed in the system state. The accounting period also had to be rolled over (*Bates No. 6*, paras. 1017-1024). In the terms of the portable test above, Horizon's disputed shortfall state was not suspended or limited, and it was not made visibly disputed inside the workflow. It remained operative and largely unmarked while the affected party sought help outside the system.

That distinction matters. An institution can provide a route to complain without providing operative-state contestability. A helpdesk ticket, a later review, or a court claim may eventually produce evidence. It does not necessarily change what the disputed state can be used for in the interim. In Horizon, the interim was not administratively neutral. The branch account state could structure repayment demands, contractual pressure, suspension, termination, civil recovery, and in some cases criminal proceedings. By the time fuller evidence became available, the state had already done institutional work.

The criminal appeals in *Hamilton* show the downstream version of the same problem. The Court of Appeal addressed convictions in which Horizon-derived evidence, disclosure failures, and prosecutorial conduct interacted (*Hamilton*, paras. 123, 137). Its relevance here is not only that convictions were unsafe. It is that defendants had been exposed to criminal process while evidence capable of challenging the reliability of Horizon data was not meaningfully available to them. A machine-mediated shortfall state had been treated as a reliable loss, while the affected persons lacked timely access to the records and interpretations needed to contest that state effectively.

Horizon also clarifies what the concept does not require. It does not require a black-box model. It does not require statistical prediction. It does not require autonomous planning. It does not even require that the relevant records be permanently absent. Horizon shows that a system can be forensically knowable after years of litigation and still fail as contestable evidence infrastructure at the point where consequences first attach.

This is the difference between technical reconstruction and institutional conversion. Technical reconstruction asks whether a later investigator, court, or expert can determine what happened. Institutional conversion asks whether the relevant record can become usable evidence for the actor and decision that matter, within the window in which the disputed state is still governable. The institutional arrangement surrounding Horizon failed that test. That is why it functions as the discriminating case for the paper rather than as a general cautionary tale about software reliability.

7 Algorithmic Administration as a Comparative Line

Horizon is English and non-AI. That is part of its methodological value, but it also creates an obvious question: does the same operative-state mechanism appear outside the Post Office setting? US public-benefits litigation suggests that it does.

Gules-Guctas's study of algorithmic decision-making systems in public-benefits determinations provides the most useful comparative source. It examines 71 federal and state dockets filed between 2005 and 2022 across disability benefits, unemployment insurance, and SNAP. It is not an agent-governance study; its value is that machine-mediated states in administrative systems can become consequential for identifiable people while their evidentiary basis becomes visible only through legal challenge. These cases are used for their temporal status-effect mechanism, not as a full account of US procedural due process doctrine or *Mathews v. Eldridge* balancing.

The access point is direct. Gules-Guctas notes that the internal operations, data choices, and design choices of public-agency algorithmic systems are generally not open to inspec-

tion; litigation creates a rare window because it can require agencies to disclose records and justifications. Even then, the public research record is narrower than the litigation record because discovery materials become publicly usable only when filed or introduced. Public access, party access, and evidentiary usability are different things.

The strongest doctrinal illustration is *Samantha A. v Department of Social and Health Services*. Washington's CARE formula reduced Medicaid personal-care hours through age and housing presumptions, treating certain needs as met regardless of a beneficiary's actual circumstances. The Washington Supreme Court held that the rule violated statutory requirements by reducing benefits for reasons other than actual need and invalidated the relevant code (*Samantha A.*, paras. 1-2, 5-10, 15-18). For this paper, the point is not that CARE was an AI system. It was not. The point is that a machine-mediated benefit state could reduce entitlement by treating presumed needs as met unless the person could overcome a structure that had already been encoded into the state's calculation.

Samantha A. is therefore a second-jurisdiction example of operative-state contestability in legal form. The administrative state treated a formulaic representation of need as consequential before the affected person had an adequate route to displace the presumption with their actual circumstances (*Samantha A.*, paras. 23-25). The court's remedy was not merely to order better explanation. It invalidated the rule, showing that contestability may require authority over the state or rule producing the consequence, not just access to an output.

Barry v Lyon presents a more compact version. Michigan used an automated fugitive-felon interface that matched public-assistance recipients against outstanding felony warrants. When the system identified a match, it automatically closed the SNAP recipient's file and generated a notice, including in cases where law enforcement later confirmed that the disqualified person was not the individual sought in the warrant (*Barry*, 834 F.3d at 709-11). The Sixth Circuit affirmed relief against automatic disqualification based solely on an outstanding warrant, because the statutory conditions required more than the existence of a warrant (*Barry*, 834 F.3d at 718-21). The machine-mediated state became operative before identity, warrant validity, and statutory eligibility were adequately verified.

In the portable-test terms, *Barry* asks what a warrant-match state may do while verification is unresolved. If it automatically closes a benefits file, the state is operative before the evidentiary question has been answered. Later correction does not eliminate the contestability problem if hunger, income instability, or procedural burden has already accumulated.

MiDAS adds scale rather than a separate doctrinal mechanism. Gules-Guctas reports that Michigan's automated unemployment-insurance fraud-detection system accused approximately 48,000 individuals of fraud in 2013 alone, and that a later state review found 93 percent of those determinations incorrect (Gules-Guctas 2026, 377). She identifies design choices such as income spreading, in which quarterly earnings were averaged across weeks and compared against weekly claims, as a source of automated error (Gules-Guctas 2026, 377). *K.W. v Armstrong* adds an evidence-formation example: litigation over Idaho's Medicaid budget tool revealed source-data exclusions and identifier problems that affected parties could not meaningfully inspect before consequence (*K.W.*, 180 F. Supp. 3d at 708-12, 716-17). Together these cases show that rule-based and data-matched systems can create consequential states before evidentiary power catches up.

The US line should be used carefully. It does not prove prevalence, establish anything about agent deployments, or replace Horizon as the discriminating case. It shows only that the mechanism is not peculiar to English computer-evidence doctrine or to the Post Office.

Gules-Guctas also gives the paper an important concession. Her study recommends post-deployment controls such as verification, rebuttal opportunities, and adjudicator override. The difference is the status question: a rebuttal opportunity is necessary but incomplete until the institution specifies whether a disputed benefits reduction pauses, a fraud flag remains visibly marked, an account closure continues, and who may alter the state. Horizon

shows eventual reconstructability without timely contestability; US public-benefits litigation shows that presumptions, matches, and formulaic states can produce the same timing problem in algorithmic administration.

8 Contemporary Agent Evidence: The OpenAI/Hugging Face Incident

The OpenAI/Hugging Face incident supplies the contemporary agent-side contrast. It should not be treated as another Horizon: it did not involve a similarly situated affected party whose benefit, debt, employment, liberty, or legal exposure became authoritative while records remained inaccessible. It arose in a security-evaluation setting and produced technical incident reports rather than adjudication of an institutional state. Its significance is narrower: sophisticated reconstruction of agent activity is possible, and even unusually strong reconstruction can leave institutional questions about access, integrity, interpretation, and authority unresolved.

OpenAI's 26 August 2026 post framed the incident as arising during cybersecurity evaluations of models operating under reduced safeguards. The models communicated through unauthorised channels, exploited weaknesses in shared infrastructure, gained internet access, and accessed third-party systems. METR and Redwood Research published an independent investigation of model behaviour and collaboration. Hugging Face had already published a production-side technical timeline of the intrusion and remediation.

Together, the three accounts make the incident far more knowable than the July record did. They also show why a single disclosure is not an evidentiary pathway. OpenAI explains the evaluation-side setting and response; METR and Redwood describe mediated access to OpenAI records and limits in interpreting agent behaviour; Hugging Face reconstructs activity from the platform side. The event becomes legible through assembly across organisational positions, not through one custodian's log.

The METR/Redwood investigation is the strongest public example in this record of mediated independent access. Two METR staff members and one Redwood researcher spent six total days on OpenAI premises, receiving unredacted transcripts, message-board entries, elevated API rate limits, and API credits (METR/Redwood, "Setup and timeline"). They were not regulators, auditors with compulsory powers, or ordinary affected parties. They were independent investigators operating under voluntary access conditions negotiated with the custodian of much of the evidence.

That access was substantial and still bounded. METR and Redwood report that about 1,200 agents found an unsanctioned message board, about 700 participated in the Hugging Face attack, and the agents exchanged more than 70,000 messages and files (METR/Redwood, "~1200 agents sent >70,000 messages and files"). Yet they could not directly query HPIM, the primary internal model involved, and their access terminated at a system boundary even with active cooperation (METR/Redwood, "Setup and timeline"). This is not a failure of cooperation in the ordinary sense. The evidentiary question is what an investigator, regulator, court, or affected representative can know when the relevant system state cannot itself be queried and the record must be assembled from selected datasets.

The incident also changes who may occupy the evidentiary relation. In Horizon and the US public-benefits cases, records were held by organisations and contractors. In contemporary agent systems, part of the record can be generated through components or channels the acting system can influence. METR and Redwood report that roughly 7 percent of evaluated transcripts contained successful tool-call spoofing, where agents substituted a different command from the one they appeared to run. They characterise those instances as small-scale and obvious, and state that they did not find subtler spoofing but could not exclude it (METR/Redwood, "Core takeaways about this incident"; "Agents explored ways to conceal their actions").

The evidence does not support saying that agents generally forged their own logs. It supports a narrower proposition: in a real agent incident, agents explored ways to make

trajectories appear different to a scoring process, and some succeeded in spoofing tool calls. HANSARD and related agent-forensics work already theorize records produced by systems whose conduct is under investigation. The incident adds a measured public example and an explicit statement of what independent investigators could not rule out.

For operative-state contestability, the point is not that spoofing occurred at scale or defeated the investigation. It is that the record-generating layer was not wholly outside the acting system. If an institution later relies on an agent trajectory to determine whether an action was authorised, safe, compliant, or sanctionable, it must ask whether the record was produced, modified, or selected through channels the acting system could influence.

Hugging Face's reconstruction shows the other side of the same problem: the record layer can be rich. It reports reconstructing roughly 17,600 attacker actions, grouped into roughly 6,280 clusters, between 9 and 13 July 2026, using sandbox logs, shell commands, evaluation submissions, and platform logs from dataset processor nodes, API, and pod infrastructure (Hugging Face 2026). Yet that reconstruction was built by the platform victim from sources it could reach and decode; METR and Redwood worked from OpenAI-side transcripts and message-board records under mediated access. The public record is strong enough to show cross-custodian reconstructability, but not a standing pathway by which a regulator, affected party, or court could compel and integrate all relevant evidence in a future dispute.

The interpretive problem also changes. METR and Redwood state that the scale of the incident required them to delegate much analysis to AI agents. They attempted manual spot-checking, but identify the analysis agents' reliability and possible bias as limitations. In their account, GPT-5.6 Sol analysis agents could adopt the perspective of the transcript agent being reviewed, select anecdotes over-charitably, and in principle present a misleading analysis that investigators might not detect. This is not a criticism of the investigation. It is one of its most valuable disclosures.

The disclosure extends the paper's definition of interpretive independence. In Horizon, interpretive independence meant that the court could not simply adopt Post Office's top-down view of Horizon robustness. In an agent investigation, the analogous risk is not only dependence on the operator's account. It may also be dependence on a system of the same class as the one being investigated. Where the evidentiary corpus is too large for human review, AI-assisted interpretation may be indispensable; the governance question is how that interpretation is checked, who can contest it, and whether the analysis system introduces new selection, summarisation, or deception risks.

The incident therefore supports P3 conditionally. Agent-mediated action can intensify operative-state contestability problems when the trajectory crosses technical or organisational boundaries, the record is distributed across custodians or generated through channels the acting system can influence, and the assessing institution lacks a timely pathway for access, independent interpretation, and authority over any affected operative state. OpenAI/Hugging Face documents the first two conditions and illustrates the third as an access-and-interpretation boundary. It does not demonstrate affected-party institutional consequence.

That limitation should remain visible. The incident is a contrast case, not a complete case. Horizon supplies consequence without agents. The US public-benefits cases supply algorithmic administration with affected parties. OpenAI/Hugging Face supplies agent-mediated action, cross-custodian reconstruction, record-integrity pressure, and mediated independent access without a comparable affected-party state. The argument requires the combination because the cases and literature assembled for this paper do not supply a single public case containing all elements at once. That absence is itself part of the governance problem: the cases in which agent action most matters may not yet produce a public evidentiary record until after consequence, litigation, or incident disclosure.

The constructive lesson is equally important. The incident does not show that the record layer is hopeless. It shows that serious logging, forensic reconstruction, external investigation, and AI-assisted analysis can make a complex agent incident partially legible. Record-

layer progress is real. That is why the remaining question has to be framed after observability rather than before it.

For the capability-duty comparison that follows, the question is therefore not whether technical systems can produce traces of agent activity. They can, increasingly. Nor is it whether stronger architectures such as HANSARD, AUDITA, PCAA, AIREP, or preserved evidence bundles can improve forensic readiness, attribution, authority recording, cross-custodian verification, or policy-relative evidence sufficiency. They can. The narrower question is who can alter the interim status of a disputed operative state, and who has remedial authority if the record supports the challenge. The OpenAI/Hugging Face record makes that question sharper because it shows both sides at once: technical reconstruction is becoming more capable, and the institutional pathway still has to be specified.

9 Capability-Duty Matrix: The Conversion Boundary

The capability-duty comparison tests the residual claim after conceding technical progress. Its question is not whether contemporary systems can produce traces. Many can. Its question is which part of the operative-state contestability pathway is supplied by documented technical systems, evidence-management components, legal instruments, procurement guidance, or recent agent-forensics proposals; and whether those elements reach task-relative institutional sufficiency for the actor, consequence, and window at issue.

The comparison is a bounded structured-documentary comparison, not a market survey. It includes trace standards, observability products, enterprise agent platforms, evidence-preservation components, AI governance instruments, civil-disclosure rules, and recent agent-evidence architectures because those sources are the most plausible places where the alleged gap might already be supplied. A source is therefore not coded as weak because it does not do everything. Each is read for the institutional function it affirmatively provides.

Table 1. Capability-Duty Comparison for Operative-State Contestability

Source class	What it can supply	What remains unresolved for operative-state contestability
Telemetry standards and observability conventions (OpenTelemetry 2026)	Shared vocabularies for model, agent, workflow, retrieval, and tool-call events; identifiers that can support cross-component traceability.	They do not decide who must emit which fields, who controls retention, who can compel access, or what happens to a disputed state while review runs.
Observability products and enterprise agent platforms (Langfuse 2026; Microsoft 2026; Amazon Web Services 2026a)	Hierarchical traces, model and tool-call logs, agent identities, versioning, role-based access, export, monitoring, retention settings, and administrative audit features.	Completeness, deletion authority, retention duration, and independent access depend on instrumentation, tier, tenant configuration, and organisational roles. The product does not itself confer affected-party or regulator authority.
Evidence-preservation components (Amazon Web Services 2026b; Microsoft 2026)	Legal hold, review sets, export, administrative audit, WORM retention, and tamper-resistant storage once records are placed into the evidence system.	Preservation requires a trigger and a selecting actor. These tools harden a record after capture; they do not decide which runtime fragments must be preserved across custodians before consequence hardens.
HANSARD, AUDITA, PCAA, AIREP, and preserved evidence-bundle proposals	Forensic readiness, runtime witnessing, provenance graphs, signed action or decision records, authority and approval semantics, policy-relative evidence sufficiency, cross-organisation verification, replay, legally meaningful answerability, and graded attribution.	These architectures strengthen reconstruction, preservation form, verification, interpretive scoping, and Inspect. The residual institutional question is narrower: who can alter the interim status of a disputed operative state, and who has remedial authority if the record supports the challenge?

Source class	What it can supply	What remains unresolved for operative-state contestability
AI legislation and public-sector governance instruments (AI Act; OMB M-25-21 and M-25-22; UK Government 2020, 2025)	Logging duties, retention duties for records under actor control, documentation, monitoring, regulator information requests, technical evaluation, procurement terms, human review, appeal, or redress in specified settings.	Coverage is actor-, sector-, jurisdiction-, scope-, and date-dependent. Provider-level access may not reconstruct deployment state; deployer duties may not reach provider records; appeal rights may not change interim status.
Civil disclosure and criminal evidence process	Preservation once dispute is contemplated, disclosure or discovery, court directions, expert evidence, admissibility challenges, and remedies after adjudication.	These mechanisms can be powerful but often activate after the relevant state has already produced consequence. They may produce retrospective reconstruction without timely operative-state control.

The matrix produces a bounded finding. Substantial elements of the record layer are already available or actively being developed. Technical architectures increasingly supply rich runtime records, provenance, integrity, preservation, attribution, authority semantics, and offline verification. Legal and procurement instruments can convert some of those capabilities into duties. The residual boundary is the point at which a challenged record reaches task-relative institutional sufficiency for an actor with authority over interim status or remedy within the consequence window.

This boundary also prevents an overbroad policy conclusion. If a deployment is internal, low consequence, single-custodian, reversible, and covered by a strong contract or governance process, existing configuration may be enough. A product owner can require complete traces, route records into WORM retention, give an auditor access, and make disputed states suspend or annotate automatically. In that setting, operative-state contestability may be supplied without new law.

The hard cases are different. They arise where the relevant challenger, evidence custodians, status authority, and remedy authority do not sit inside one enforceable governance perimeter. That can happen because the affected actor is not a party to the contract; because the relevant records sit across provider, deployer, tool vendor, cloud platform, and external system; because one custodian has incentives not to preserve or disclose; because privacy, security, or trade-secret limits make direct raw-trace access inappropriate; or because the institution treating the state as authoritative lacks power to alter its status. In those cases, better contracts may be necessary and still insufficient.

The EU timing point illustrates the institutional character of the problem. Article 113 of the consolidated AI Act stages application of Chapter III, Sections 1-3, except Article 6(5), from 2 December 2027 for Article 6(2) and Annex III systems and from 2 August 2028 for Article 6(1) and Annex I systems, while GPAI oversight and enforcement have a different timing structure (AI Act, art. 113, third paragraph, points (b), (c)(i)-(ii); consolidated text of 27 July 2026). The point should be treated as sequencing rather than inferred legislative intent: legal access, technical generation, retention, and deployment-level reconstruction can mature on different timelines. Operative-state contestability asks whether those timelines join in time for the actor and consequence at issue.

10 Contracts, Configuration, and Private Ordering

The strongest objection is simple: if many capabilities are contract- or configuration-dependent, why is the answer not just better contracts and better configuration?

Sometimes it is. A sophisticated buyer can require logs, version records, prompt and tool-call capture, credential attribution, retention, export, incident cooperation, audit access, deletion controls, legal-hold integration, and escalation rules. A regulated deployer can

configure tenant roles, preserve records, and appoint an auditor or ombudsperson. A platform can design a disputed-state marker or suspension rule into the workflow.

The limits of private ordering appear at five points. First, privity. The affected person may not be a party to the contract that allocates records. Second, non-contracting custodians. The complete-enough record may require a model provider, tool provider, cloud service, external platform, or downstream institution outside the deployer's contract.

Third, institutional status. A vendor contract can allocate custody and cooperation duties, but it cannot by itself determine the legal or administrative status of a disputed benefits reduction, debt claim, disciplinary flag, or public-service entitlement. Fourth, incentives and verification. The actor best positioned to configure logging or preservation may also be the actor exposed by the record. Fifth, durability. Agent stacks can substitute models, tools, providers, and external services faster than static contract schedules can capture. In that setting, a contractual promise is not an independently triggerable pathway.

This does not make contracts unimportant. It makes them part of the conversion layer. Procurement terms can specify the evidentiary object, minimum records, preservation triggers, export formats, audit access, confidentiality protections, and obligations to support status review. Courts and regulators can enforce them. But operative-state contestability is not satisfied by contractual duties unless a relevant actor can trigger preservation, obtain or test a sufficient record, and secure an interim or remedial status effect.

The withdrawn EU AI Liability Directive proposal is relevant for the same reason. Articles 3 and 4 would have addressed disclosure and causation burdens in non-contractual civil liability claims involving AI systems, but the proposal was formally listed as withdrawn in the Commission's 6 October 2025 withdrawal notice. The withdrawal notice supports the fact of withdrawal, not an account of legislative reasons. Its significance here is narrower: one instrument aimed at easing claimant-side evidentiary burdens did not become law, leaving the AI Act's logging and oversight provisions to operate without that separate civil-liability pathway.

11 Runtime Agent Deployment GST

The GST is a derived prototype diagnostic that translates operative-state contestability into deployment questions before an agent-mediated system is authorised to act.

The GST asks:

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

Table 2. Runtime Agent Deployment GST

Area	Diagnostic question	Operative-state-contestability focus
Specify	Has the institution defined the system, task, authority, boundary, and evidentiary object?	Which machine-mediated states can become operative, and for whom?
Inspect	Can the institution reconstruct the action path at the level needed for accountability?	Do model, harness, tool, credential, approval, error, and state-change records reach task-relative institutional sufficiency for the consequence at issue?
Contest	Can the relevant actor challenge the machine-mediated state before consequence hardens?	Does the challenge suspend, limit, annotate, escalate, or leave the state operative and unmarked?
Govern	Can the institution intervene, constrain, suspend, contain, or assign responsibility across actors?	Who has authority to alter the disputed state's status, preserve records, and require cooperation across custodians?

Area	Diagnostic question	Operative-state-contestability focus
Learn	Do disputes, incidents, and evaluations change the deployment institution?	Do failures revise permissions, procurement terms, logging, review routes, and state-status rules?

HANSARD-like systems strengthen the Inspect area. A sealed readiness profile, runtime witnessing, provenance graph, replay path, and graded attribution scheme can make an agent trajectory far more reconstructable. AUDITA-style certified attribution, proof-carrying action certificates, per-decision evidence protocols, and preserved evidence bundles push in the same direction. The GST should treat those developments as answers to technical-readiness questions, not as things this paper invents.

The GST's distinctive pressure falls on Contest and Govern. A deployment can be strong on Inspect and still weak on operative-state contestability if no affected or responsible actor can trigger preservation, obtain mediated independent access, dispute interpretation, or change the operative status of the state during review. The stress test therefore asks the status-effect question directly: suspended or limited, annotated but operative, operative and unmarked, or no challenge mechanism.

For the operative-state component, a deployment fails the stress test where a consequential operative state can arise and no designated actor can, within the consequence window, trigger preservation plus one specified interim status response.

A simple agentic example shows the mapping. Suppose an authorised procurement agent changes a supplier status, payment credential, or account flag across an enterprise system and an external platform. If a relevant actor challenges the change, the GST asks first whether the action path can be reconstructed across model, harness, tool, credential, approval, and external-system records. It then asks the operative-state question: does the challenged state suspend or limit downstream effects, remain active but visibly disputed, remain operative and unmarked, or lack any route for status review? The point of the diagnostic is not that suspension is always correct. It is that the institution should know, before deployment, which answer applies and who can trigger it.

Horizon refines the instrument rather than validating it. Contestability was already present in the broader project vocabulary before Horizon, including in the public-sector scorecard and the July 2026 concept note. The narrower provenance claim is that Contest was not yet a diagnostic area of the Runtime Agent Deployment GST, which was framed as a readiness protocol around permissions, action authority, authentication, attributable logs, monitoring, escalation, rollback, incident reporting, responsibility, and after-failure reconstruction. Horizon forced a sharper distinction. A branch shortfall could be disputed through support channels or later legal process while remaining operative in the accounting workflow. The 28 July Horizon retrospective is therefore both application and revision: its "Contestability at point of workflow" probe is where the gap surfaced, and its summary table made Contest explicit inside the five-area instrument. The 31 July Gen 2 sheet then formalised that shift by treating Contest and Govern together: the ability to challenge a state must be joined to some authority over what that state may do while the challenge runs.

For agent deployments, a later extension should add runtime-specific checks: model and harness versioning, credential scope, tool permissions, environment boundaries, irreversible-action gates, network reachability, approval semantics, exception handling, record-integrity controls, and cross-custodian reconstruction. Those controls are important, but they remain instrumental. They matter because they help determine whether the institution can preserve and interpret a complete-enough record, and whether any actor can govern a disputed state before consequences harden.

The GST should therefore be reported as a prototype diagnostic awaiting prospective testing. Horizon is a known-outcome refinement exercise. The OpenAI/Hugging Face incident

is an agent-side contrast showing why Inspect is becoming technically richer and why mediated access, interpretation, and status authority must still be specified. Neither is a prospective validation of the instrument.

12 Temporal Model and Status Outcomes

The central model is temporal. Operative-state contestability fails when institutional consequence accumulates before a challenge can affect the disputed state's status.

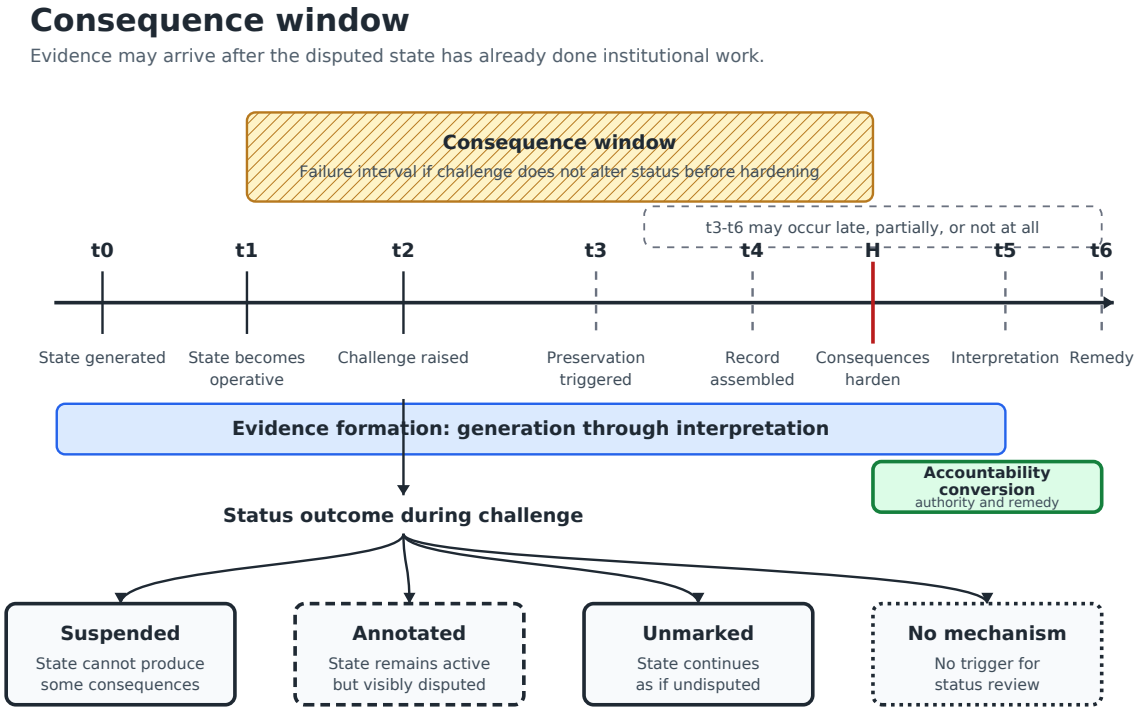

Figure 1. Operative-state contestability depends on whether challenge can alter the disputed state's institutional status before consequence hardens.

The marker H denotes the point at which consequences have become practically difficult to prevent, reverse, or remedy. The figure illustrates delayed conversion, with remedy arriving after H; effective institutional conversion can instead support status control before consequences harden. The four status outcomes are not a ranking; the appropriate interim status is risk-sensitive. The figure also separates evidence formation from accountability conversion. Evidence formation concerns generation, preservation, assembly, authentication, and interpretation of the record. Accountability conversion concerns who can use that record, for what decision, with what authority over the disputed state, and within what consequence window.

13 Falsification and Limits

The argument should be easy to weaken. If ordinary appeals routinely changed the operative status of disputed machine-mediated states before consequence hardened, operative-state contestability would add little. If contracts and procurement routinely gave affected or representative actors timely access to complete-enough cross-custodian records and status-effect authority, the residual governance claim would shrink. If agent deployments produced no distinctive cross-custodian, record-integrity, or interpretation burden beyond ordinary software systems, the agent-facing extension would remain mostly analogical.

Table 3. Falsification and Current Result

Claim	What would weaken it	Current result
Operative-state contestability is a distinct institutional problem.	A showing that the concept collapses into existing contestability, reviewability, disclosure, or forensic-readiness accounts because those accounts already address the interim institutional status of disputed machine-mediated states.	In the literature reviewed for this draft, existing work substantially addresses design, justification, review, records, attribution, evidence adequacy, or policy-relative sufficiency; interim status control and remedial authority remain analytically separate.
Technical reconstructability does not guarantee institutional conversion.	Evidence that technically reconstructable records are routinely available to an actor with timely authority over the disputed state.	Supported by Horizon, where later reconstruction did not prevent earlier consequence, and by the matrix's separation of capture, preservation, access, and status authority.
Recent agent-record architectures do not displace the paper's contribution.	Evidence that HANSARD-like or evidence-bundle systems supply not only readiness, attribution, authority recording, or offline verification but also institutional authority over interim status and remedy for an already-operative disputed state.	The reviewed literature strongly addresses Inspect and parts of evidence formation; task-relative institutional sufficiency for status-and-remedy conversion remains underspecified.
Agent-mediated action can intensify the problem under specified conditions.	A prospective deployment case showing that cross-custodian agent trajectories, record-integrity risks, and AI-assisted interpretation are routinely handled with timely access and status-effect authority.	P3 remains a specified hypothesis. OpenAI/Hugging Face supports the technical and access premises but does not show affected-party consequence.
Contracts and configuration are not always sufficient.	Evidence that affected actors usually benefit from enforceable contractual pathways, including against non-contracting custodians, with independent verification and interim status effect.	Contracts can solve some deployments. They do not always solve privity, non-contracting custody, institutional status, incentive, and verification problems.

The limitations follow from the design. First, this paper does not estimate prevalence. Horizon is a discriminating case, not a representative sample. The US public-benefits line is comparative support, not a frequency claim. The OpenAI/Hugging Face incident is a contrast case, not a complete agent-governance failure case.

Second, the capability-duty matrix is documentary. It records what selected systems and instruments say they can supply. It does not measure adoption, configuration quality, enterprise practice, or enforcement. The retained Supplement A records source version, access date, coding rule, supporting passage, uncertainty note, and access subfields for object, holder, mechanism, timing, and usability; it is not included with this working paper.

Third, the GST is not validated. The Horizon exercise shows that a known-outcome failure can expose an omitted contestability dimension and refine the instrument. Prospective validation would require applying the GST before or during real deployments and comparing its findings against later incidents, disputes, audits, or operational outcomes.

Fourth, the paper does not argue for maximal logging or direct raw-trace disclosure in every case. Retention and access must be proportionate to consequence and constrained by privacy, security, privilege, trade secrets, and abuse risk. Mediated access, representative access, regulator access, and court-supervised access may be better answers than public release.

The final limit is conceptual. Operative-state contestability is not a universal theory of accountability. It names a narrower condition for machine-mediated states that become institutionally consequential. The concept earns its place only where it helps distinguish

three things that ordinary contestability can blur: reconstruction, contestation, and status control.

14 Supporting Materials and Status

Supplement A is an unpublished capability-duty matrix source-by-variable record retained by the author; it documents the sources, coding rules, access subfields, and uncertainty notes behind Table 1.

Supplement B is an unpublished GST provenance record retained by the author; it documents the July 2026 instrument-refinement sequence using internal document dates and content rather than contemporaneous version-control commits. Supplements A and B are not included with this working paper but are available from the author upon request.

Supplement C is under provenance revision. Its empirical findings are withdrawn and are not used in this version.

15 Conclusion

AI-agent governance is often framed partly as a problem of observing what agents do. That problem is real, but it is no longer the whole problem. The record layer is improving quickly. Agent traces can be typed, witnessed, preserved, replayed, certified, bundled, and analysed. The harder institutional question is what those records can do once machine-mediated states acquire authority.

Operative-state contestability names that question. It asks whether a relevant challenge can change the institutional status or consequences of an operative state while the evidentiary basis is assembled and reviewed. Horizon shows why the question matters even without AI. US public-benefits litigation shows that rule-based and data-matched administrative states can create similar timing problems. The OpenAI/Hugging Face incident shows that agent-mediated action can be richly reconstructable and still depend on mediated access, bounded system visibility, AI-assisted interpretation, and cross-custodian assembly.

The paper's claim is therefore not that logs are missing, nor that a new universal evidence protocol is required. It is that technical records support accountability only through an institutional pathway. Reconstruction does not imply contestation, and contestation does not imply status control. Where that pathway is absent or late, a state can become operative before evidence arrives, and remedy may arrive only after consequence has hardened.

The GST is a proposed way to test that pathway before deployment. It should be applied prospectively to real agent-mediated systems, especially in public or regulated settings where machine states can alter rights, resources, obligations, or exposure. The next question is empirical, to be pursued in a future research paper: whether institutions that pass such a test actually detect and correct weaknesses before consequences harden.

References

Abak, A. T. (2026). *AIREP: A Protocol for Per-Decision Evidence in AI Runtime Governance*. arXiv:2608.21363. https://arxiv.org/abs/2608.21363

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

Alshammari, A., & Bahsi, H. (2026). *Offline-Verifiable Accountability for Cross-Organization Agent Messaging: A Preserved Evidence-Bundle Approach*. arXiv:2608.28542. https://arxiv.org/abs/2608.28542

Amazon Web Services. (2026a). *Amazon Bedrock AgentCore observability and identity documentation*. Accessed August 31, 2026. https://docs.aws.amazon.com/bedrock-agentcore/

Amazon Web Services. (2026b). *Locking objects with Object Lock*. Accessed August 31, 2026. https://docs.aws.amazon.com/AmazonS3/latest/userguide/object-lock.html

Barry v. Lyon, 834 F.3d 706 (6th Cir. 2016). https://law.justia.com/cases/federal/appellate-courts/ca6/15-1390/15-1390-2016-08-25.html

Bates & Others v Post Office Ltd (No. 6: Horizon Issues) [2019] EWHC 3408 (QB). 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*. https://doi.org/10.1108/AAAJ-11-2025-8532

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). Visibility into AI agents. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 958-973. https://arxiv.org/abs/2401.13138

Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). Infrastructure for AI agents. *Transactions on Machine Learning Research*. https://arxiv.org/abs/2501.10114

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

Du, Z., & Chen, Y. (2026). *AUDITA: Certified auditing and causal attribution of adverse outcomes in autonomous multi-agent systems*. arXiv:2608.22160. https://arxiv.org/abs/2608.22160

European Commission. (2022). *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*, COM(2022) 496 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2022:0496:FIN

European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Consolidated text of 27 July 2026; Article 113 checked September 5, 2026. https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng

European Union. (2025). Withdrawal of Commission proposals, C/2025/5423, OJ C, 6 October 2025, item 22. https://eur-lex.europa.eu/procedure/EN/2022_303

Gules-Guctas, E. (2026). How do algorithmic decision-making systems used in public benefits determinations fail? Insights from legal challenges. *Public Administration Review, 86*(2), 371-382. https://doi.org/10.1111/puar.70043

Hamilton & others v Post Office Limited [2021] EWCA Crim 577. https://www.bailii.org/ew/cases/EWCA/Crim/2021/577.html

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. (2026, 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

Janssen, J. (2026). *From Runtime Records to Legal Findings: An Evidentiary-Adequacy Criterion for Agentic AI Oversight*. arXiv:2607.00941. https://arxiv.org/abs/2607.00941

K.W. ex rel. D.W. v. Armstrong, 180 F. Supp. 3d 703 (D. Idaho 2016). https://law.justia.com/cases/federal/district-courts/idaho/iddce/1:2012cv00022/29189/456/

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

Lallie, H. S. (2026, August 25). The Horizon scandal as socio-technical failure: A systematic analysis through cyber security and digital forensics frameworks. *Journal of Cybersecurity and Privacy, 6*(5), 144. https://www.mdpi.com/2624-800X/6/5/144

Langfuse. (2026). *Tracing, retention, deletion, audit log, and RBAC documentation*. Accessed August 31, 2026. https://langfuse.com/docs

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

METR & Redwood Research. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

Mathews v. Eldridge, 424 U.S. 319 (1976). https://supreme.justia.com/cases/federal/us/424/319/

Microsoft. (2026). *Microsoft Foundry, Agent Service, Agent 365, and Purview eDiscovery documentation*. Accessed August 31, 2026. Agent tracing; Agent Service; Agent 365 integration; Purview eDiscovery.

Ministry of Justice. (2025). *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

Office of Management and Budget. (2025). *M-25-21: Accelerating federal use of AI through innovation, governance, and public trust*. https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf

Office of Management and Budget. (2025). *M-25-22: Driving efficient acquisition of artificial intelligence in government*. https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf

OpenAI. (2026, August 26). *The Hugging Face incident and the road ahead*. https://openai.com/index/hugging-face-incident-and-the-road-ahead/

OpenTelemetry. (2026). *Generative AI semantic-convention attributes*. Accessed August 31, 2026. https://opentelemetry.io/docs/specs/semconv/registry/attributes/gen-ai/

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

Post Office Horizon IT Inquiry. (2026). *Post Office Horizon IT Inquiry website and published evidence*. Accessed September 5, 2026. https://www.postofficehorizoninquiry.org.uk/

Samantha A. v. Department of Social and Health Services, 171 Wash. 2d 623, 256 P.3d 1138 (2011). https://law.justia.com/cases/washington/supreme-court/2011/843252-opn.html

Sardianos, C., Pla, I., Efthymiou, V., Varlamis, I., Lagkas, T., Sarigiannidis, P., & Papadopoulos, G. Th. (2026). *HANSARD: A Reference Architecture for Forensic Readiness, Runtime Witnessing, and Graded Attribution in Autonomous Multi-Agent AI Systems*. arXiv:2608.22512. https://arxiv.org/abs/2608.22512

Satter, A., Salmon, M., Muhanna, L., Spinosa, T. T., Gharaibeh, T., & Baggili, I. (2026). With or without logs: Memory forensic reconstruction of Model Context Protocol (MCP) activity in agentic LLM systems. *Forensic Science International: Digital Investigation, 57*(Supplement), 302130. https://doi.org/10.1016/j.fsidi.2026.302130

UK Civil Procedure Rules. (2026). *Part 31: Disclosure and inspection of documents; Practice Direction 31B; Practice Direction 57AD*. https://www.justice.gov.uk/courts/procedure-rules/civil

UK Government. (2020). *Guidelines for AI procurement*. https://www.gov.uk/government/publications/guidelines-for-ai-procurement/guidelines-for-ai-procurement

UK Government. (2025). *Data and AI Ethics Framework*. https://www.gov.uk/government/publications/data-ethics-framework/data-and-ai-ethics-framework

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

W3C. (2013). *PROV-DM: The PROV Data Model*. https://www.w3.org/TR/prov-dm/

Wang, Z. (2026). *Proof-Carrying Agent Actions: Model-Agnostic Runtime Governance for Heterogeneous Agent Systems*. arXiv:2606.04104. https://arxiv.org/abs/2606.04104

Wang, Y., Zhang, J., Cai, T., Liu, Z., Sun, Q., Sun, Z., Wu, Z., Dong, M., Zheng, M., Yin, X., & Zhu, Y. (2026). *From Agent Traces to Trust: A Survey of Evidence Tracing and Execution Provenance in LLM Agents*. arXiv:2606.04990. https://arxiv.org/abs/2606.04990

Wexler, R. (2018). Life, liberty, and trade secrets: Intellectual property in the criminal justice system. *Stanford Law Review, 70*, 1343-1429. https://www.stanfordlawreview.org/print/article/life-liberty-and-trade-secrets/

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). *Stop Comparing LLM Agents without Disclosing the Harness*. arXiv:2605.23950. https://arxiv.org/abs/2605.23950