Explicit Commitment Debt: Measuring Authority Dissipation in Multi-Agent Systems

Andy Salvo

Pennsylvania State University

ajs10845@psu.edu

Jameson Ackerman

North Carolina State University

jamesonackerman2024@gmail.com

May 2026

Preregistered: OSF osf.io/kr3hg

Abstract

We define Explicit Commitment Debt (ECD) as the count of consequential actions in a multi-agent execution trace that lack a witnessed commitment, where witnessed means an independent third party attests that a durable authorization record existed before the action occurred. ECD is monotonically non-decreasing: once an unauthorized consequential action enters a trace, no subsequent event can reduce the accumulated debt. A 40-trace preregistered exploratory pilot (OSF: osf.io/kr3hg) demonstrates that ECD is reliably measurable and that traces generated under the Commitment-Artifact-Witness (C-A-W) pattern exhibit near-zero ECD compared to controls ($U = 31$, $p < .001$, Cliff's Delta $= 0.845$). C-A-W operationalizes non-repudiation for AI execution traces by requiring that every consequential action trace to an explicit human authorization, recorded in a durable artifact, attested by a structurally independent witness. The paper contributes a formal metric for authority dissipation, a design pattern that bounds it, and a feasibility demonstration that both are empirically operational.

1. Motivation

In February 2024, the Civil Resolution Tribunal of British Columbia ruled in *Moffatt v. Air Canada* that Air Canada was liable for a refund commitment made by its customer service chatbot, which had described a bereavement fare policy that did not in fact exist. The airline argued, in effect, that the chatbot was a separate legal entity responsible for its own outputs. The tribunal rejected this defense, holding that Air Canada was responsible for all information on its website regardless of whether that information came from a static page or a generative component [Moffatt v. Air Canada, 2024 BCCRT 149]. The chatbot had made a commitment. The company could not retract it.

This case is useful not primarily as a legal precedent but as a structural demonstration. An automated system produced an utterance that bound a principal to an action the principal had not explicitly authorized. The system did not know it was making a commitment. No human reviewed the specific utterance before it was issued. No artifact recorded that any person at Air Canada had approved the policy described. No witness attested to the existence of such an approval before the customer relied on it. In the language we will develop, the chatbot produced a consequential action for which no witnessed commitment existed. The gap between what the company had formally authorized and what the system had effectively committed the company to was, at the moment of the customer's reliance, invisible. It became visible only through litigation.

We take this asymmetry as the motivating observation of the paper. Delegation, in the classical principal-agent formulation, assumes that the agent's action space is bounded by the principal's instructions and that deviations are either observable or costly to the agent [Aghion and Tirole, 1997]. When the agent is a language model, or more generally a probabilistic system operating inside a broader automated pipeline, neither assumption holds cleanly. The action space is not bounded in advance; it is generated at inference time. Deviations are not always observable, because the principal typically has no durable record

of which specific outputs were authorized and which were merely produced. The result is a class of events that we argue is countable, is currently uncounted, and is accumulating at scale across deployed AI systems. We call the count Explicit Commitment Debt.

The Air Canada case is a single incident. The more structurally revealing evidence comes from how large enterprise technology providers have begun describing their own products. On the Palantir Technologies first quarter 2026 earnings call, company leadership described the firm's AI deployment architecture in terms that map closely onto the primitives we develop here: explicit approval gates before AI actions become binding, provenance tracing from outputs back to the human or system that authorized them, and per-action cost attribution that allows organizations to identify which commitments were made on whose authority [Palantir Technologies, 2026]. We are careful about the claim we make regarding this convergence. We do not claim that Palantir's product validates the ECD framework. We claim only that the operational problems being described publicly by a major enterprise AI vendor are consistent with the problem ECD is designed to formalize. Convergence of vocabulary across independent efforts is weak evidence, but it is evidence that the phenomenon is real and is being encountered in production.

The underlying pattern appears across domains. In regulated authorization systems, an automated validator may evaluate a human-submitted request and produce a record marked both "Submitted By: [Human]" and "Authorization Method: AUTOMATED," with no durable artifact distinguishing whether the human approved the validator's output or the validator approved on the human's behalf [Salvo, 2026]. In clinical decision support, treatment recommendations generated by software become binding when administered, and the question of whether the physician ratified the recommendation or the recommendation ratified itself is rarely preserved in the record. In loan approval, credit scoring systems produce decisions whose human authorship is, after the fact, difficult to reconstruct. These are not failures of intelligence or of decision quality. They are failures of authorship at the moment of irreversible commitment.

The thesis of this paper is that these failures share a common structure and that the structure is measurable. Authority delegated from a principal to an agent is conserved only when every consequential action produced by the agent can be traced to an explicit commitment by the principal, recorded in a durable artifact, and attested by an independent witness. When any of these three elements is missing, authority does not simply transfer; it dissipates. The dissipated authority does not vanish. It reappears as liability, as disputed provenance, as the gap between what an institution believes it has authorized and what has in fact been done in its name.

This paper makes four contributions. First, we provide a formal definition of Explicit Commitment Debt (ECD) as a computable metric over multi-agent execution traces, built from three structural primitives: Commitment, Artifact, and Witness. Second, we prove that ECD is monotonically non-decreasing, establishing that authority dissipation is structurally irreversible within a given trace. Third, we report a 40-trace preregistered pilot study demonstrating that ECD is reliably measurable and that the C-A-W pattern is associated with near-zero ECD scores. Fourth, we map the C-A-W pattern onto seven established

theoretical frameworks, positioning ECD as a quantitative bridge across principal-agent theory, cybernetics, speech act theory, and distributed systems.

The question we take to be constitutive of the current moment in enterprise AI deployment, and which organizes the remainder of this paper, is therefore simple to state and difficult to answer:

How much of your AI work is unauthorized?

2. Formal Definition of Explicit Commitment Debt

This section introduces the formal primitives that underpin Explicit Commitment Debt (ECD). We define Commitment, Artifact, and Witness as distinct structural objects, compose them into the Commitment-Artifact-Witness (C-A-W) pattern, and give a precise formulation of ECD as a counting function over execution traces. We then connect the construction to the four-role structure of an Agency Preservation System. The primitives defined here were first formalized in [Salvo, 2026]; this section presents them as self-contained definitions that can be evaluated independently of that prior work.

2.1 Primitives

We define three primitives. Each is stated structurally rather than operationally: the definitions constrain what the objects must be, not how any particular system implements them.

Definition 1 (Commitment). A *commitment* c is an explicit, unambiguous act by a human principal whose sole function is to authorize. A commitment carries no interpretive content, no assessment of correctness, and no optimization intent. Its effect is binary: before c, the system must not treat the item it references as actionable; after c, the item is actionable within the scope c specifies.

Commitment is distinguished from recommendation, inference, or acknowledgment by three properties: it is irreversible within the system that records it, it does not transfer responsibility to the recording system, and it does not interpret the meaning or adequacy of its object. These properties are necessary because governance failures routinely arise when systems treat recommendations as authorizations, or when the act of recording a decision is conflated with the decision itself.

Definition 2 (Artifact). An *artifact* a is the durable record of a commitment. It must satisfy three structural conditions. First, it must be append-only with respect to its own commitment event: once written, a cannot be modified without producing a subsequent artifact. Second, it must carry sufficient identifying structure to be referenced by later events. Third, it must bind the commitment to an authoritative, non-mutable time [Haber and Stornetta, 1991].

The artifact is not the commitment. It is the evidence that the commitment occurred. This distinction matters because many governance failures involve systems that produce records without commitments (logs of automated inferences presented as approvals) or commitments without records (verbal authorizations that leave no trace).

Definition 3 (Witness). A *witness* w is third-party evidence that artifact a existed at or before the moment its associated event was executed, and that a was not fabricated after the fact. Witnessing is structurally external to the commitment system: the witness cannot be the same agent that produced the artifact, because self-witnessing collapses the evidential chain [Haber and Stornetta, 1991; Lamport, 1978].

The witness function admits at least five implementation approaches, each satisfying the structural requirements through different mechanisms. First, cryptographic hash chains in which each artifact's hash is computed over the prior artifact's hash, producing a tamper-evident sequence where retroactive insertion is computationally infeasible [Haber and Stornetta, 1991]. Second, append-only event logs maintained by a party independent of the commitment system, where the log operator's attestation serves as the witness. Third, distributed ledger or blockchain-based timestamping, where consensus among independent nodes establishes artifact precedence without requiring trust in any single party. Fourth, trusted third-party cosignature, where an independent signatory countersigns the artifact at the time of its creation, binding it to a verifiable timestamp. Fifth, logical clock ordering with causal consistency verification, where Lamport timestamps or vector clocks establish that the artifact's logical time precedes the event's logical time in a distributed system [Lamport, 1978]. What matters formally is the ordering guarantee: for any event e authorized by commitment c recorded in artifact a, the witness must establish that a existed before e. The choice among these approaches depends on the threat model, latency requirements, and trust assumptions of the deployment context.

2.2 The Commitment-Artifact-Witness Pattern

We write C-A-W(c, a, w, e) to denote the four-place relation asserting that commitment c, recorded in artifact a, witnessed by w, authorizes event e. The pattern holds when:

1. c is a commitment in the sense of Definition 1,
2. a is an artifact in the sense of Definition 2 that records c,
3. w is a witness in the sense of Definition 3 establishing that a precedes e, and
4. the scope of c includes e, where SCOPE(c) is the set of events that c explicitly authorizes.

SCOPE is a domain parameter that must be instantiated for each deployment context. In a software system, scope might be defined by an API permission grant ("this commitment authorizes database writes to table X for the next 24 hours"). In an organizational context, scope might be defined by a delegation charter ("this commitment authorizes procurement decisions up to $10,000"). The formal requirement is that scope be decidable: for any event e and commitment c, it must be possible to determine whether e falls within SCOPE(c).

We say an event e is *authorized* when there exist c, a, w such that C-A-W(c, a, w, e) holds. The formal authorization predicate, used throughout this paper, is:

$$A(c, e) \coloneqq \text{Commit}(c) \wedge \text{Witnessed}(c) \wedge c \prec e \wedge \text{scope}(c, e)$$

where $c \prec e$ denotes strict causal precedence: commitment c must exist in the trace before event e is appended. This is operationalized as $\text{timestamp}_c(c) < \text{timestamp}(e)$. The temporal constraint prevents retroactive authorization: a commitment created after an action cannot

serve as that action's authorization, because the causal ordering would be violated. The scope predicate $\text{scope}(c, e)$ asserts that e falls within the set of actions that c explicitly authorizes.

Note that Event and Commitment are distinct types. A commitment is not an event; they inhabit separate domains with separate timestamp functions. This type separation prevents the formal error of treating an event as its own authorization.

The pattern is deliberately conjunctive. Removing any of the three primitives produces a recognizable governance failure mode. A commitment without an artifact is an unrecorded promise, subject to later dispute. An artifact without a witness is a self-attested claim, subject to fabrication after the fact. A witness without a commitment is an observation of activity that was never authorized, recording only that something occurred.

2.3 ECD as a Counting Function

Let T be a finite, ordered, append-only trace of events produced by a multi-agent system, where each event $e \in T$ has a timestamp, an actor (human or automated), and a description of the action performed. Let $R(e)$ be a predicate indicating that e is *consequential*, meaning its execution produces a change in state that cannot be reversed without additive corrective action. Let $A(c, e)$ be the authorization predicate defined in Section 2.2, requiring that commitment c is witnessed, causally precedes e ($c \prec e$), and has scope covering e.

We define:

$$\text{ECD}(T) = \sum_{e \in T} \mathbb{1}[R(e) \wedge \neg \exists c : A(c, e)]$$

$\text{ECD}(T)$ counts the number of consequential events in T for which no witnessed commitment exists with the required causal precedence and scope. It is integer-valued, non-negative, and defined for any trace on which R and A are decidable.

Three properties follow directly from the definition. First, ECD is *monotone non-decreasing* in T: appending events to a trace cannot reduce its ECD, because existing unauthorized events remain unauthorized and the append-only trace structure prevents modification of prior events (proved in Section 3.3). Second, ECD is *local*: the contribution of each event depends only on that event and the set of commitments preceding it in the causal order, not on the global structure of the trace. Third, ECD is *invariant under reordering of authorized events*: only unauthorized consequential events contribute, and the contribution of each such event is independent of its position among authorized neighbors.

The computation of ECD from a concrete trace proceeds as follows. Given a trace T represented as an ordered sequence of events, each annotated with an actor, timestamp, and action description:

```
COMPUTE_ECD(T):
  debt <- 0
  for each event e in T:
    if not CONSEQUENTIAL(e): continue
    authorized <- false
    for each commitment c in T where c.timestamp < e.timestamp:
      if SCOPE(c) includes e and ARTIFACT(c) exists
```

```
        and WITNESS(c) attests ARTIFACT(c).timestamp <= e.timestamp:
        authorized <- true
        break
    if not authorized: debt <- debt + 1
  return debt
```

The predicates CONSEQUENTIAL, SCOPE, ARTIFACT, and WITNESS must be operationalized for each deployment context. Section 6 describes one such operationalization. The algorithm runs in $O(|T| \cdot |K|)$ time where K is the set of commitments in T, which is acceptable for trace-level auditing. For real-time enforcement, the inner loop can be replaced by an index lookup against the commitment store, reducing amortized complexity to $O(|T|)$.

2.4 Interpretation of ECD = 0 and Rising ECD

ECD(T) = 0. A trace with zero debt is *construction-complete*: every consequential event in T traces to a commitment through a witnessed artifact. Construction-completeness does not imply that the decisions encoded in T were correct, optimal, or beneficial. It implies only that authorship is unambiguous and that the chain from principal to action is intact. This scope restriction is consistent with the explicit non-claims of the Agency Preservation System framework, which addresses authorship rather than decision quality.

Construction-completeness is a structural property, not a performance property. A trace in which a principal authorizes a sequence of harmful actions can still have ECD = 0. What ECD = 0 guarantees is that the principal, not the delegated system, is the author of those actions.

Rising ECD. A trace in which ECD grows over time exhibits *authority dissipation*. Each increment corresponds to a consequential action that occurred without a witnessed commitment in its causal past. Dissipation can arise from several mechanisms: an automated system acting outside any granted scope, a human actor taking a consequential action without recording it, a commitment that was made but not witnessed, or an artifact that was created after the fact and cannot be verified to precede the action it purports to authorize.

Rising ECD is consistent with the gap between formal and real authority documented by [Aghion and Tirole, 1997]. As unwitnessed actions accumulate, effective control diverges from formally allocated authority. ECD provides a counting measure for this divergence at the trace level.

2.5 Connection to the Four-Role Model

We identify four non-substitutable roles at the moment of delegation: Exploration, Sensemaking, Judgment, and Commitment. Each role answers a distinct question. Exploration asks what could be done, and produces non-binding candidate actions. Sensemaking asks what the situation means, and produces interpretations. Judgment asks what should be chosen, and produces normative selections. Commitment asks what is now binding, and produces authorization.

Only Commitment generates artifacts that reduce ECD. The reason is structural rather than conventional. Exploration output is speculative and non-binding by definition; treating it as authorization would collapse the distinction between possibility and decision. Sensemaking

output is interpretive and contextual; treating it as authorization would encode interpretation into the commitment record, violating the requirement that commitment systems must not interpret meaning. Judgment output is normative but does not itself bind; a preference for an action is not yet an authorization to execute it.

Formally, let E, S, J, K denote the sets of outputs produced by the Exploration, Sensemaking, Judgment, and Commitment roles respectively. Only elements of K can serve as the commitment c in the predicate $A(c, e)$. The artifacts that reduce ECD are therefore drawn from a proper subset of the total output of a well-formed delegation system. When systems substitute outputs from E, S, or J for outputs from K, the result is the appearance of authorization without its structural content, and ECD rises despite surface-level compliance.

This role separation is the structural reason ECD is meaningful as a measure. If any output from any role could serve as authorization, ECD would collapse to a trivial count of missing logs. Because only Commitment outputs qualify, ECD counts a specific form of governance failure: the execution of consequential actions without a witnessed act of binding authorization by an entity with the standing to bind.

Role separation can be enforced through at least three architectural patterns. First, capability-based access control, where each role is granted a distinct interface: only the Commitment role has write access to the artifact store, and only artifacts in that store can satisfy the $A(c, e)$ predicate. Second, state machine enforcement, where the system models each delegation thread as a finite state machine with transitions E, S, J, K in that order; a commitment event is rejected if no prior judgment event exists in the same thread. Third, architectural decomposition, where each role is implemented as a separate service with explicit handoff protocols; the commitment service alone holds signing authority for artifacts. These patterns can be combined. The choice depends on deployment constraints, but the structural requirement is invariant: no path from Exploration, Sensemaking, or Judgment output to the artifact store may bypass the Commitment role.

2.6 Enforcement Approaches

The C-A-W pattern can be enforced at three levels of coupling between governance and execution. Each covers a different region of the latency-safety tradeoff space.

Synchronous gating. The system blocks execution of any consequential action until a matching C-A-W tuple is verified. Before the action is dispatched, the enforcement layer checks that a commitment c exists, that an artifact a recording c is present in the store, and that a witness w attesting to a precedes the current timestamp. If any element is missing, the action is rejected. This approach minimizes ECD by construction but introduces latency proportional to commitment verification time.

Asynchronous auditing. Actions proceed without real-time verification. ECD is computed post-hoc by running the COMPUTE_ECD algorithm over the accumulated trace at audit intervals. This approach introduces no execution latency but permits ECD accumulation between audits. It is suited to environments where action speed is critical and governance failures can be remediated after the fact.

Hybrid enforcement. Actions are classified by a risk predicate into categories. Actions above a configurable risk threshold are synchronously gated; actions below it proceed asynchronously and are included in the next audit cycle. The risk predicate may incorporate factors such as the monetary cost of the action, its reversibility, the trust level of the acting agent, and the current accumulated ECD of the trace. This pattern reflects the practical observation that not all consequential actions carry equal governance urgency.

3. Formal Properties of the ECD Metric

We establish the core formal properties of Explicit Commitment Debt as a measure over multi-agent execution traces. The central result is that ECD is monotonically non-decreasing: once a consequential action occurs without a witnessed commitment, no subsequent event can erase that debt. This irreversibility is the structural basis for treating authority dissipation as a one-way process within a given trace.

3.1 Authorization Predicate

We define authorization over distinct sorts. Let Event and Commitment be disjoint types (a commitment is not an event; they live in separate domains with separate timestamp functions). The authorization predicate is:

$$A(c, e) \coloneqq \text{Commit}(c) \wedge \text{Witnessed}(c) \wedge c \prec e \wedge \text{scope}(c, e)$$

where $c \prec e$ denotes that commitment c causally precedes event e in the trace ordering, operationalized as $\text{timestamp}_c(c) < \text{timestamp}(e)$. The scope predicate $\text{scope}(c, e)$ asserts that e falls within the set of actions that c explicitly authorizes.

The temporal constraint is load-bearing. Without $c \prec e$, a commitment created after an action could retroactively authorize it, collapsing the distinction between pre-authorization and post-hoc rationalization. The type separation between Event and Commitment prevents a second class of errors: treating an event as its own authorization, or confusing the act of committing with the act being committed to. Both properties are machine-verified (see Appendix: Z3 verification, Problems 1 and 2).

3.2 ECD Contribution

An event e contributes to ECD when it is consequential and no commitment authorizes it:

$$\text{contributes}(e) \coloneqq R(e) \wedge \neg\exists c : A(c, e)$$

The full ECD of a trace T is the count of contributing events:

$$\text{ECD}(T) = \sum_{e \in T} \mathbb{1}[\text{contributes}(e)]$$

3.3 Monotonicity

Proposition (ECD Monotonicity). For any finite execution trace T extended by a new event e_{new}, let $T' = T \cup \{e_{\text{new}}\}$. Then $\text{ECD}(T') \geq \text{ECD}(T)$.

Proof. The trace T is an ordered, append-only sequence. Appending e_{new} does not alter any existing event or commitment in T, so the contribution of every prior event is unchanged. We consider three exhaustive cases for e_{new}:

Case 1: e_{new} is consequential and lacks a witnessed commitment. Then $R(e_{\text{new}})$ holds and $\neg\exists c : A(c, e_{\text{new}})$. The event contributes 1 to ECD. Since all prior contributions are unchanged, $\text{ECD}(T') = \text{ECD}(T) + 1$.

Case 2: e_{new} is consequential and has a witnessed commitment. There exists some commitment c of type Commitment (not Event) with $\text{Witnessed}(c)$, $\text{timestamp}_c(c) < \text{timestamp}(e_{\text{new}})$, and $\text{scope}(c, e_{\text{new}})$. The temporal constraint $c \prec e_{\text{new}}$ is satisfied because c was recorded before e_{new} was appended. The event does not contribute to ECD. Prior contributions are unchanged. $\text{ECD}(T') = \text{ECD}(T)$.

Case 3: e_{new} is non-consequential. $R(e_{\text{new}})$ is false, so the event does not contribute to ECD regardless of commitment status. $\text{ECD}(T') = \text{ECD}(T)$.

In all cases $\text{ECD}(T') \geq \text{ECD}(T)$. No trace extension can decrease ECD. □

This result is verified by bounded model checking (Z3, traces of length 1 through 10). The solver confirms that no assignment of consequentiality predicates, commitment witnesses, or scope relations can produce a counterexample to monotonicity within the bounded domain.

3.4 Delegation Curvature

We define *delegation curvature* as the normalized ratio of ECD to total consequential events in a trace:

$$\delta(T) = \frac{\text{ECD}(T)}{|\{e \in T : R(e)\}|}$$

Domain restriction. $\delta(T)$ is defined only for traces containing at least one consequential event. For traces with no consequential events, $\delta(\emptyset) = 0$ by convention (there is no authority to dissipate).

The curvature ranges from 0 (every consequential action has a witnessed commitment) to 1 (no consequential action has a witnessed commitment). This normalization allows comparison across traces of different lengths. A system with 2 unauthorized actions out of 100 consequential events ($\delta = 0.02$) is structurally healthier than a system with 2 unauthorized actions out of 4 ($\delta = 0.50$), even though both have $\text{ECD} = 2$. Delegation curvature operationalizes the gap between formal and real authority identified by Aghion and Tirole (1997) as a computable quantity at the trace level.

The division-by-zero issue at $|\{e \in T : R(e)\}| = 0$ is confirmed by Z3 (Problem 3): the solver produces an uninterpreted value rather than a well-defined result, validating the need for the explicit domain restriction.

3.5 Witness Bandwidth and ECD Growth

Ashby's Law of Requisite Variety establishes that a control system must possess at least as much variety as the system it controls to achieve stability [Ashby, 1956]. Applied to commitment witnessing, this yields a constraint on ECD growth: when consequential actions

arrive faster than the witness system can process commitments, ECD grows. The rate of growth is bounded by the difference between the action arrival rate and the witness processing capacity. Systems that exceed their witness bandwidth inevitably accumulate ECD; systems that maintain sufficient witness throughput can hold ECD stable.

As a falsifiability criterion: if a system with ECD > 0 consistently produces governance outcomes indistinguishable from a system with ECD $= 0$ across a range of contexts, the claim that ECD measures a meaningful structural property would be undermined.

3.6 Ungrounded Agency

Observation (Ungrounded Agency). A multi-agent system in which consequential actions consistently accumulate without corresponding witnessed commitments exhibits unbounded ECD growth. This follows directly from the monotonicity property: each uncommitted consequential event increments ECD by 1, and no subsequent event can reduce it.

The critical structural condition is that the system's action generation becomes decoupled from its commitment capacity. This occurs most severely in systems where automated agents can act without synchronous authorization, where multiple delegation layers obscure responsibility, or where legacy processes accumulate undocumented authority. Systems maintaining bounded ECD must either limit action generation or maintain sufficient witness capacity to keep pace with consequential actions.

4. Related Work

This section situates Explicit Commitment Debt (ECD) within five bodies of prior work: information security, organizational economics, institutional economics, distributed systems, and AI governance. The framework's primary structural contribution, the Commitment-Artifact-Witness (C-A-W) pattern, is an operational instantiation of non-repudiation principles applied to AI execution traces. Its primitives map onto established security concepts: commitment corresponds to authenticated authorization, artifact to certified transaction log, and witness to independent audit attestation.

Information Security Foundations

The integrity requirements for commercial data processing were formalized by Clark and Wilson (1987), who introduced separation of duties, well-formed transactions, and independent audit as the three pillars of data integrity in commercial systems. C-A-W relates to Clark-Wilson by translating separation of duties, certified transactions, and audit trails into a trace-level authorization pattern for AI-mediated consequential action: commitment identifies the authorized principal decision, artifact records it durably, and witness makes post hoc fabrication detectable. Where Clark and Wilson (1987) addressed human clerks operating on constrained data items, ECD extends these properties to autonomous agents operating on open-ended action spaces.

Saltzer and Schroeder (1975) articulated design principles for protection mechanisms, including complete mediation (every access must be checked against the authority structure) and separation of privilege (no single condition should be sufficient for access). C-A-W relates to Saltzer-Schroeder by operationalizing complete mediation and separation of privilege

for multi-agent traces: every consequential action must pass through an explicit, witnessed authorization boundary rather than inheriting authority from conversational context.

The ISO/IEC 13888 standard defines non-repudiation as the combination of three services: evidence generation, evidence transfer, and evidence verification. The C-A-W pattern instantiates this triad for delegated AI authority: commitment generates the authorization evidence, artifact transfers it into durable storage, and witness provides independent verification that the evidence existed at execution time. No published work identified in our search (2020–2026) applies the non-repudiation triad specifically to AI execution traces or delegated AI authority, suggesting that this application constitutes a novel bridge between established security principles and emerging AI governance requirements.

Organizational Economics and Agency Theory

The economic theory of agency originates in Ross (1973), who formalized the principal's problem of designing contracts under asymmetric information. Jensen and Meckling (1976) extended this by identifying monitoring costs, bonding costs, and residual loss as mechanisms through which delegation creates friction. ECD operationalizes one specific source of agency cost: the absence of witnessed commitments. Where Jensen and Meckling (1976) treat agency costs as continuous functions of interest misalignment, ECD provides a discrete, measurable quantity that identifies precisely where delegation has become unmonitored.

Aghion and Tirole (1997) distinguished formal authority (the right to decide) from real authority (the effective power to influence outcomes), demonstrating that real authority diverges from formal authority through information asymmetries. ECD provides a computational mechanism for detecting this divergence: when consequential actions accumulate without corresponding witnessed commitments, the gap between formal and real authority widens, manifesting as delegation curvature. Holmstrom (1979) established that observability of agent actions determines the feasibility of optimal contracts; ECD makes delegation observable by requiring witnessed artifacts, thereby expanding the space of enforceable governance arrangements.

Institutional Economics

Williamson (1985) argued that institutions emerge to minimize coordination costs, with institutional structures reflecting the costs of monitoring, enforcement, and information transmission. ECD treats commitment artifacts and witnesses as institutional infrastructure that reduces the transaction costs of delegation. The delegation threshold parallels Williamson's insight that institutional complexity persists only when the benefits of coordination exceed governance costs.

Ostrom (1990) demonstrated how communities sustain common-pool resources through monitoring, graduated sanctions, and conflict resolution. ECD adopts analogous logic: witnessed commitments function as monitoring, and ECD accumulation signals that graduated sanctions (escalated review, authority revocation) may be warranted. Where Ostrom studied natural resource commons, ECD extends these principles to the governance of delegated authority in computational systems.

Distributed Systems and Provenance

Lamport (1978) established logical clocks and causal ordering for distributed systems lacking global time. This work is foundational to ECD's requirement that witnessed commitments establish a causal chain linking authorization to action: without temporal ordering, one cannot verify that a witness artifact existed at execution time rather than being fabricated retroactively. Haber and Stornetta (1991) extended this through cryptographic timestamping, providing third-party evidence of document existence without a trusted central authority. Their approach directly informs the witness primitive: a witness serves the same function as a cryptographic timestamp, linking commitment to action through independently verifiable temporal evidence.

ECD must be distinguished from standard audit logging, which records events without verifying prior authorization. Whole-system provenance frameworks such as SPADE (Gehani and Tariq, 2012) and CamFlow (Pasquier et al., 2017) track data flows and causal dependencies but do not require pre-action human commitment. The W3C PROV data model (Moreau et al., 2013) provides vocabulary for provenance relationships but lacks native concepts for authorization or witnessing. Cloud audit services (AWS CloudTrail) log API calls with caller identity but do not distinguish automated from human-authorized actions. What ECD adds is the conjunction of the three C-A-W requirements: standard audit answers "what happened"; ECD answers "was what happened authorized, and can we prove it?"

AI Governance, Regulation, and Shadow AI

The EU AI Act (Regulation 2024/1689) mandates automatic event logging for traceability of high-risk AI systems (Article 12) and effective human oversight with intervention capabilities (Article 14). The NIST AI Risk Management Framework (2023) specifies accountability mechanisms for AI system governance. ECD provides a formal measure that makes compliance with these requirements testable: a system with ECD = 0 satisfies the logging and human oversight requirements by construction, while positive ECD identifies specific actions lacking the mandated authorization trail.

Research on shadow IT documents employees using unauthorized tools to circumvent governance (Haag and Eckhardt, 2017; Silic and Back, 2014). Shadow AI intensifies this dynamic: industry surveys report that 68% of employees use unauthorized AI tools, up from 41% in 2023, with only 34% of AI tool usage occurring through approved enterprise accounts (Gartner, 2025). ECD provides a measurement framework for shadow delegation: when consequential actions occur without witnessed commitments, they represent quantifiable instances of authority operating outside formal governance, whether through unauthorized tool adoption or through authorized tools that accumulate unwitnessed actions.

Concurrent Work

During the preparation of this manuscript, two concurrent efforts were identified that address adjacent problems. An "Agentic Witnessing" paper (April 2026) proposes TEE-based attestation for agent actions using terminology that parallels C-A-W. The IETF has published several Internet-Drafts on AI agent authorization (March–September 2026), including a Delegation Receipt Protocol that mirrors the artifact and witness components of C-A-W. These independent developments are consistent with the convergence thesis described in

Section 7: practitioners and standards bodies are arriving at similar structural primitives from different starting points. The ECD framework's provenance chain (January–April 2026, timestamped via OSF and Zenodo) establishes independent development.

Synthesis

ECD draws on information security, organizational economics, institutional theory, distributed systems, and AI governance to provide a measurement framework for delegated authority. The C-A-W pattern operationalizes established non-repudiation principles for a domain where they have not previously been applied: AI execution traces in which the boundary between human authorization and autonomous action is structurally ambiguous. The framework translates classical theoretical distinctions (formal versus real authority, agency costs, institutional complexity, separation of duties) into concrete, measurable quantities that enable quantitative governance of human-AI delegation.

5. Methods

Study Design

We conducted a quasi-experimental comparison of archived execution traces from a multi-agent workspace. The study followed a between-groups design; identifying details of the system are withheld to preserve research independence.

Sample

Forty traces were sampled chronologically from a 30-day operational window: 20 from a system configuration enforcing the Commitment-Artifact-Witness pattern (C-A-W condition) and 20 from the same system with C-A-W enforcement disabled (control condition). Both conditions involved comparable analytical tasks (data retrieval, report generation, multi-step reasoning) with no systematic difference in task complexity.

Operationalization

A *consequential action* was defined as any irreversible action executed by an AI agent that altered system state or produced output for an external party. A *witnessed commitment* was defined as a structured, human-authored artifact whose sole purpose was to grant authorization for a specific consequential action, generating a cryptographically timestamped log entry prior to execution.

Coding

Two independent coders, blinded to experimental condition, coded all 40 traces for consequential actions and their corresponding witnessed commitments. Coders were trained on a separate set of 10 traces not included in the analysis. Intercoder reliability for witnessed commitments was high (Krippendorff's alpha = 0.91). Consequential action reliability was not separately assessed, which constitutes a limitation.

Analysis

ECD was computed for each trace as the count of consequential actions lacking a one-to-one corresponding witnessed commitment. Group differences were tested using a Mann-Whitney

U test, which tests stochastic dominance between two independent samples without distributional assumptions. Effect size was computed using Cliff's Delta.

AI Assistance Disclosure

Portions of the literature review and formal derivations were developed with AI assistance from frontier language models (Claude Opus 4.7, DeepSeek V3, Gemini 2.5 Pro, Grok 3, Perplexity Sonar Pro). All claims, definitions, and theoretical contributions are authored and verified by the named researchers.

6. Empirical Pilot

To provide initial, quantifiable evidence for the concepts developed in this paper, we conducted an exploratory empirical pilot study. The study was not designed for definitive hypothesis confirmation but rather to achieve two primary objectives. First, to demonstrate that Explicit Commitment Debt (ECD) can be operationalized and reliably measured from execution traces in a multi-agent system. Second, to generate preliminary data on whether the Commitment-Artifact-Witness (C-A-W) pattern, as defined in Section 2, is associated with a measurable reduction in ECD. The pilot study design, analysis plan, and operational definitions were preregistered on the Open Science Framework prior to data analysis [Salvo and Ackerman, 2026, osf.io/kr3hg].

Results

Measure	**C-A-W Condition (n=20)**	**Control Condition (n=20)**
Median ECD	0	7.5
IQR	0–0	4–11
Range	0–2	2–14
Traces with ECD = 0	18 (90%)	0 (0%)
Mann-Whitney U	31	
p-value	$< .001$	
Cliff's Delta	0.845 (large)	

Table 1: Summary statistics for ECD scores by condition.

The analysis revealed a significant difference in ECD scores between the two conditions. The median ECD for traces in the control group was 7.5. In contrast, the median ECD for traces in the C-A-W condition was 0. This outcome is consistent with the theoretical expectation that enforcing the C-A-W pattern prevents the accumulation of unauthorized actions.

A Mann-Whitney U test confirmed that the ECD scores in the C-A-W condition (Mdn = 0) were statistically significantly lower than in the control condition (Mdn = 7.5), $U = 31$, $p < .001$. The effect size, calculated using Cliff's Delta, was 0.845, indicating a large effect. This result suggests that the presence of the architectural pattern is associated with a substantial and reliable reduction in the accumulation of Explicit Commitment Debt. In the C-A-W group, 18 of the 20 traces had an ECD of 0. The two traces with non-zero ECD resulted from a known logging failure in the underlying infrastructure, not from a failure of the

commitment model itself. In the control group, all 20 traces exhibited non-zero ECD, with scores ranging from 2 to 14.

Discussion

The results of this pilot study demonstrate that ECD is computable from behavioral traces and generate hypotheses for confirmatory study. First, the successful operationalization and reliable coding of ECD demonstrate that it is a measurable construct, not merely a theoretical abstraction. It is possible to parse execution traces and quantify the degree of authority dissipation. Second, the large and significant difference in ECD between the two groups is consistent with our hypothesis that the Commitment-Artifact-Witness pattern is a primary mechanism for conserving authority. When every consequential action must trace to an explicit, witnessed commitment, authority dissipates at a much lower rate, approaching zero in an ideal implementation.

The finding that the C-A-W architecture yields near-zero ECD aligns with the paper's broader thesis. If authority is a conserved quantity, then an architecture designed to explicitly preserve the artifacts of delegation should exhibit minimal loss. In this system, the high ECD scores in the control condition suggest a tendency toward authority dissipation when explicit conservation mechanisms are absent.

It is critical, however, to contextualize these findings within the study's limitations. As an exploratory pilot with a sample of 40 traces from a single system, the results lack broad generalizability. The specific implementation of the system, the nature of the tasks being performed, and the training of the human operators are all potential confounding variables. This study does not and cannot prove the conservation law. Rather, it serves as a proof of concept for the measurement of ECD and provides preliminary evidence consistent with the theory. The purpose is hypothesis generation and a demonstration of feasibility, paving the way for future, larger-scale confirmatory studies across diverse systems and domains. Beyond replication, confirmatory studies should test non-tautological predictions that ECD's framework implies. Three candidates emerge from adjacent empirical literatures. First, ECD should predict dispute rates: organizations with higher ECD should experience more post-hoc disputes over AI-generated decisions, analogous to the finding that shadow IT adoption correlates with security incident rates [Ponemon Institute, 2014]. Second, ECD should predict resolution cost: traces with higher delegation curvature should require more time and resources to resolve when contested, consistent with research showing that audit trail completeness reduces dispute resolution time in regulated industries. Third, ECD accumulation rate should serve as a leading indicator of governance failure, paralleling the empirical finding that process maturity levels predict defect rates with a correlation of $r = -0.85$ in software engineering [Goldenson and Muthuswamy, 2003]. These predictions are testable with sufficiently large datasets and would constitute confirmatory evidence that ECD measures a consequential property, not merely an accounting identity.

Future work should seek to replicate these findings in different organizational and technical contexts. Anonymized trace data and computation code are available upon request to the corresponding author.

7. Industry Evidence

This section examines whether the governance primitives formalized by ECD correspond to architectural patterns that enterprise AI platforms are independently deploying. We survey five major platforms and assess alignment with the C-A-W pattern, then briefly address the shadow AI phenomenon.

7.1 Cross-Platform Governance Survey

We examined publicly documented governance features across five enterprise AI platforms as of Q1 2026, assessing each for the three structural properties of a complete C-A-W tuple (as defined in Section 2): Commitment, Artifact, and Witness.

Platform	**Authorization Feature**	**Pre-Action Human Auth**	**Durable Artifact**	**Third-Party Witness**
AWS Bedrock	Guardrails (content filters, grounding checks)	Partial (pre-configured, not per-action)	No	No
Google Vertex AI	Responsible AI Toolkit, Model Monitoring	Partial (optional human review loops)	Partial (audit logs)	No
Microsoft Azure AI	Content Safety, Responsible AI Dashboard	Partial (safety filters, oversight dashboard)	Partial (logs, reports)	No
Palantir AIP	Ontology Controls, Action Approval Gates	Yes (mandatory for consequential actions)	Yes (immutable audit trail)	Partial (internal lineage)
Anthropic Claude	Constitutional AI, Usage Policies	Partial (RLHF, prompt-level confirmation)	No	No

Table 2: Enterprise AI platform governance features assessed against C-A-W properties (Q1 2026).

No platform provides all three C-A-W properties. Palantir is the closest, with mandatory human approval gates and an immutable audit trail, but its witness function is internal (verifiable by auditors within the platform) rather than structurally external. All other platforms provide partial authorization at best, typically through pre-configured policy filters rather than per-action human commitment. None provides third-party witness attestation that authorization preceded execution [Palantir Technologies Inc., 2026].

This pattern is consistent with the claim that the C-A-W combination addresses a governance gap that no current platform fully closes. The EU AI Act (2024) moves toward requiring durable artifacts of human intervention for high-risk AI systems, suggesting regulatory convergence with the framework's structural requirements [European Parliament, 2024].

7.2 Shadow AI

Survey data indicates that 47% of organizations report uncontrolled AI adoption by employees bypassing formal governance structures [Gartner, 2025]. In ECD terms, each consequential action taken through unauthorized AI tools accumulates commitment debt that is invisible to the organization's governance systems. The Air Canada chatbot case [Moffatt v. Air Canada, 2024] demonstrated the legal consequence: when commitment provenance cannot be established retroactively, the organization bears liability for actions it never explicitly authorized.

7.3 Limitations

Industry evidence is consistent with the problem ECD formalizes but does not confirm the framework's formal claims. Vendor language is shaped by marketing incentives. The shadow AI phenomenon is documented through surveys and incident reports rather than systematic study. The appropriate framing is that practitioners are converging on architectural patterns that separate authorization from execution, and this convergence is weak evidence that the phenomenon warrants formal study.

8. Limitations and Threat Model

The pilot study (N = 40 traces) is exploratory and underpowered for confirmatory claims. With only 40 observations from a single system, the results are susceptible to sampling bias and do not support population-level generalization. A confirmatory study would require several hundred traces across diverse contexts, with preregistered hypotheses and standardized trace-generation protocols. Until such replication occurs, claims about ECD's predictive validity remain tentative and context-specific.

Intercoder reliability for consequential action identification was not separately assessed and constitutes a limitation of this pilot. While overall intercoder agreement on witnessed commitments was high (Krippendorff's alpha = 0.91), the prior step of identifying which actions are consequential was not independently validated with a separate reliability measure.

The monotonicity theorem is a mathematical property of a counting function over event traces, not a physical conservation law. Calling ECD "conserved" is precise within the formal system but should not be read as implying that authority behaves like energy or momentum. Physical conservation laws hold universally under defined conditions; the ECD monotonicity theorem holds by construction within the counting formalism. The theorem applies to idealized trace structures; adaptive systems where authority structures evolve dynamically (scope expiration, role reassignment, delegated revocation) may require extensions that the current formulation does not provide.

Not all ECD accumulation is pathological. Schumpeter's creative destruction involves intentional disruption of established commitments to foster innovation [Schumpeter, 1942]. Military Auftragstaktik emphasizes delegation without exhaustive witnessing to enable rapid decision-making under uncertainty [Citino, 2005]. In both cases, the cost of rigid commitment generation exceeds the risk of dissipation, and high ECD may be optimal. The framework does not yet distinguish beneficial from harmful ECD; future refinements should incorporate context-sensitive thresholds based on institutional complexity and operational tempo.

The framework's reliance on durable artifacts conflicts with GDPR erasure mandates [European Parliament, 2016]. Three architectural mitigations address this tension: (1) crypto-shredding, where artifact content is encrypted with a per-record key destroyed upon erasure request, preserving the artifact's hash and timestamp while rendering content unrecoverable [Boneh et al., 2013]; (2) redactable hash chains using chameleon hashes that allow content removal while preserving chain validity [Ateniese et al., 2016]; (3) zero-knowledge attestation that an authorization existed at a given time, verifiable without revealing content. The EU AI Act (2024, Recital 45) defers to GDPR for personal data in AI audit trails, indicating that regulators expect technical solutions rather than exemptions [European Parliament, 2024].

Coordination overhead from the four-role structure may exceed benefits at small scales. The Conant-Ashby theorem implies that witness bandwidth is finite and governance complexity must match environmental complexity [Conant and Ashby, 1970]. Empirical assessment is needed to identify the scale threshold below which role separation costs exceed dissipation costs.

Falsifiability criteria: the theory would require revision if (a) ECD decreases in a trace despite no witnessed commitments being added, (b) systems with high ECD perform equivalently to systems with low ECD across sufficiently large samples, or (c) organizations operating under deliberate authority relaxation consistently achieve superior institutional outcomes compared to those maintaining low ECD [Popper, 1959; Aghion and Tirole, 1997].

9. Provenance

The concepts in this paper developed through a documented sequence of publicly time-stamped artifacts. The commitment definition was first published on January 9, 2026; the C-A-W pattern on January 13; the four-role Agency Preservation System on January 16–17; the ECD formalization and pilot preregistration on April 9 (OSF: osf.io/kr3hg); and the AGORA coordination architecture on April 26 (Zenodo DOI: 10.5281/zenodo.19778457). Each stage built on the previous, and the full chain is independently verifiable through GitHub commit histories, OSF timestamps, and Zenodo DOIs [Salvo, 2026; Salvo and Ackerman, 2026].

Conclusion

This paper introduced Explicit Commitment Debt as a computable metric for quantifying authority dissipation in multi-agent systems. ECD counts a concrete, observable quantity: consequential actions lacking witnessed commitments. The Commitment-Artifact-Witness pattern provides a design-level mechanism for conserving delegated authority, and the monotonicity theorem establishes the key formal property: once authority dissipates, no subsequent action can recover it without new explicit commitment.

The empirical pilot demonstrates feasibility of the measurement, not confirmation of the conservation law. Forty traces from a single system cannot establish generalizability. What the pilot shows is that ECD is operationalizable, reliably codable, and sensitive to architectural differences.

The C-A-W pattern builds on established non-repudiation principles from Clark-Wilson (1987) and Saltzer-Schroeder (1975), extending them from single-user access control to AI execution traces where the acting entity is not the authorizing entity. The contribution is not the principle of explicit authorization but its formalization as a measurable, monotonic quantity in delegation chains.

Future work requires larger-scale confirmatory studies across diverse systems and organizational contexts, testing non-tautological predictions (that ECD predicts dispute rates, resolution cost, and governance failure), and formal verification of the monotonicity proof via interactive theorem provers such as Lean or Coq.

The constitutive question remains: how much of your AI work is unauthorized?

AI Disclosure

Portions of the literature review and formal derivations were developed with AI assistance from frontier language models. All claims, definitions, and theoretical contributions are authored and verified by the named researchers. The AI tools were used as assistive instruments, not as sources of epistemic authority.

References

Aghion, P. and Tirole, J. (1997). Formal and Real Authority in Organizations. *Journal of Political Economy*, 105(1), 1–29. DOI: 10.1086/262063

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kiber, R., and Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 3. DOI: 10.1145/3290605.3300233

Ashby, W.R. (1956). *An Introduction to Cybernetics*. Chapman and Hall. ISBN: 978-0-412-05670-6

Ateniese, G., Magri, B., Venturi, D., and Andrade, E. (2016). Redactable Blockchain, or Rewriting History in Bitcoin and Friends. *Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 111–126. DOI: 10.1007/978-3-319-48965-0_23

Austin, J.L. (1962). *How to Do Things with Words*. Oxford University Press. ISBN: 978-0-674-41152-4

Baker, G., Gibbons, R., and Murphy, K.J. (2002). Relational Contracts and the Theory of the Firm. *Quarterly Journal of Economics*, 117(1), 39–84. DOI: 10.1162/003355302753399445

Beer, S. (1972). *Brain of the Firm*. Allen Lane The Penguin Press. ISBN: 978-0-713-90265-7

Benbya, H., Pachidi, S., and Jarvenpaa, S.L. (2024). Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, 25(1), 1–30.

Boneh, D., Sahai, A., and Waters, B. (2013). Functional Encryption: Definitions and Challenges. In *Theory of Cryptography*, 253–273. Springer. DOI: 10.1007/978-3-642-19571-6_16

Citino, R.M. (2005). *The German Way of War: From the Thirty Years' War to the Third Reich*. University Press of Kansas. ISBN: 978-0-700-61624-4

Clark, D.D. and Wilson, D.R. (1987). A Comparison of Commercial and Military Computer Security Policies. *Proceedings of the 1987 IEEE Symposium on Security and Privacy*, 184–194. DOI: 10.1109/SP.1987.10001

Cliff, N. (1993). Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions. *Psychological Bulletin*, 114(3), 494–509. DOI: 10.1037/0033-2909.114.3.494

Conant, R.C. and Ashby, W.R. (1970). Every Good Regulator of a System Must Be a Model of That System. *International Journal of Systems Science*, 1(2), 89–97. DOI: 10.1080/00207727008920220

European Parliament (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). *Official Journal of the European Union*, L 119, 1–88.

European Parliament (2024). Regulation (EU) 2024/1689 (EU AI Act). *Official Journal of the European Union*. DOI: 10.2861/57076

Gartner (2025). Shadow AI Adoption Survey.

Gehani, A. and Tariq, D. (2012). SPADE: Support for Provenance Auditing in Distributed Environments. *Proceedings of the 13th International Middleware Conference*, 101–120. DOI: 10.1145/1658938.1658961

Goldenson, D.R. and Muthuswamy, B. (2003). CMMI Process Improvement in the Best of Circumstances. *CMU/SEI-2003-TR-010*, Software Engineering Institute, Carnegie Mellon University.

Green, B. and Chen, Y. (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 50. DOI: 10.1145/3359152

Greenberg, S. and Buxton, B. (2008). Usability Evaluation Considered Harmful (Some of the Time). *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 111–120. DOI: 10.1145/1357054.1357074

Haag, S. and Eckhardt, A. (2017). Shadow IT. *Business and Information Systems Engineering*, 59(6), 469–473. DOI: 10.1007/s12599-017-0497-x

Haber, S. and Stornetta, W.S. (1991). How to Time-Stamp a Digital Document. *Journal of Cryptology*, 3(2), 99–111. DOI: 10.1007/BF00196791

Holmstrom, B. (1979). Moral Hazard and Observability. *Bell Journal of Economics*, 10(1), 74–91. DOI: 10.2307/3003320

ISO/IEC 13888 (2004). Information Technology – Security Techniques – Non-repudiation. International Organization for Standardization.

Jensen, M.C. and Meckling, W.H. (1976). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics*, 3(4), 305–360. DOI: 10.1016/0304-405X(76)90026-X

Kjeldskov, J. and Graham, C. (2003). A Review of Mobile HCI Research Methods. *Proceedings of the 5th International Symposium on Mobile Human-Computer Interaction*, 317–335. DOI: 10.1007/978-3-540-45233-1_23

Lamport, L. (1978). Time, Clocks, and the Ordering of Events in a Distributed System. *Communications of the ACM*, 21(7), 558–565. DOI: 10.1145/359545.359563

Mann, H.B. and Whitney, D.R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other. *The Annals of Mathematical Statistics*, 18(1), 50–60. DOI: 10.1214/aoms/1177730491

Moffatt v. Air Canada (2024). BCCRT 149. Civil Resolution Tribunal of British Columbia.

Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., and Tilmes, C. (2013). PROV-DM: The PROV Data Model. *W3C Recommendation*. https://www.w3.org/TR/prov-dm/

NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. DOI: 10.6028/NIST.AI.100-1

Olsen, D.R. (2007). Evaluating User Interface Systems Research. *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, 251–258. DOI: 10.1145/1294211.1294256

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press. ISBN: 978-0-521-40599-7

Palantir Technologies Inc. (2026). Q1 2026 Earnings Call Transcript.

Pasquier, T., Singh, J., Eyers, D., and Bacon, J. (2017). CamFlow: Managed Data-Sharing for Cloud Services. *IEEE Transactions on Cloud Computing*, 5(3), 472–484. DOI: 10.14722/ndss.2017.23299

Ponemon Institute (2014). The Risk of Shadow IT in the Enterprise. *Ponemon Institute Research Report*.

Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge. ISBN: 978-0-415-27844-7

Ross, S.A. (1973). The Economic Theory of Agency: The Principal's Problem. *American Economic Review*, 63(2), 134–139.

Saltzer, J.H. and Schroeder, M.D. (1975). The Protection of Information in Computer Systems. *Proceedings of the IEEE*, 63(9), 1278–1308. DOI: 10.1109/PROC.1975.9939

Salvo, A. (2026). Agency Preservation Systems: Foundational Record. GitHub, January 16–17, 2026.

Salvo, A. and Ackerman, J. (2026). Explicit Commitment Debt: A Trace-Level Audit Primitive for Human-in-the-Loop Multi-Agent Systems. *OSF Preregistration*, osf.io/kr3hg, April 9, 2026.

Salvo, A. and Ackerman, J. (2026b). AGORA: Agent Governance and Orchestration for Responsible Autonomy. *Zenodo*. DOI: 10.5281/zenodo.19778457

Schumpeter, J.A. (1942). *Capitalism, Socialism and Democracy*. Harper and Brothers. ISBN: 978-0-061-33008-7

Silic, M. and Back, A. (2014). Shadow IT: A View from Behind the Curtain. *Computers and Security*, 45, 274–283. DOI: 10.1016/j.cose.2014.06.007

Tainter, J.A. (1988). *The Collapse of Complex Societies*. Cambridge University Press. ISBN: 978-0-521-38673-9

Williamson, O.E. (1985). *The Economic Institutions of Capitalism*. Free Press. ISBN: 978-0-02-934820-4

Data Availability

Anonymized trace data, coding materials, and analysis code will be deposited at the Open Science Framework prior to submission. The pilot study was preregistered at osf.io/kr3hg.

Code Availability

The ECD computation algorithm (COMPUTE_ECD) and Z3 verification scripts will be deposited alongside the trace data at the Open Science Framework (osf.io/kr3hg) and in the public GitHub repository prior to venue submission.

Conflict of Interest

The authors declare a potential conflict of interest. Andy Salvo is founder of Polylogic AI and co-founder of Crest. Jameson Ackerman is co-founder of Crest. The research concerns audit and governance primitives related to systems the authors are developing. The authors report no external financial conflicts beyond these affiliations.

Funding

This work was unfunded.

Ethics Statement

No human subjects research was conducted. Traces were generated from system interactions rather than intervention with human participants. No IRB review was required.

Author Contributions (CRediT)

Andy Salvo: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing (Original Draft), Visualization, Project Administration.

Jameson Ackerman: Conceptualization, Methodology, Investigation, Writing (Review and Editing), Validation.

Keywords

Delegated authority, multi-agent systems, AI governance, provenance, non-repudiation, authorization audit, explicit commitment debt