Governing Agentic AI in FinTech

Henry Han
Data Science and Artificial Intelligence Innovation Laboratory
Baylor University, USA
Henry_Han@Baylor.edu

August 2026

Abstract

Financial institutions are beginning to delegate consequential decisions to *agentic AI* systems that decompose goals, coordinate models and tools, and act with little human oversight. However, agentic AI governance in FinTech remains under-investigated. We argue that the binding governance constraint on this shift is not capability but *verifiability*. We define the *Verifiability Gap* as the positive shortfall between the verification demanded by delegated authority and the explainability and reproducibility retained after a specific decision. The construct asks whether the system preserves evidence sufficient for a named verifier, evidentiary standard, and audit lag. We develop an agentic AI multilevel governance theory and examine its mechanisms in three studies over nine model versions, ranging from a three-billion-parameter local model to a commercial frontier system. Study 1 shows that provider releases alter historical financial actions, and that the controls replay requires belong to the provider, not the firm: the current frontier model rejects `temperature`, `top_p`, and `top_k` outright and never exposes a random seed. Under the tightest controls each endpoint allows, a local model reproduced 320 of 320 executions, while hosted models reproduced 319 of 320 and 959 of 960. Study 2 shows that orchestration is a latent policy layer. Architecture changes final actions, and no execution record repeated in any configuration at any model scale. The frontier model reproduces its own actions more often than the local ones, but its execution record is no better, and it loses a comparable share of its verdict differentiation. Capability buys a higher starting point, not auditability. Study 3 shows two deterministic credit-model versions that each reproduce their current action perfectly, while the current one cannot recover a historical one. We conceptualize reproducibility as a governance profile rather than a scalar, yielding *evidence-contingent delegation*: authority is defensible only while retained evidence can substantiate its exercise. Beyond financial services, this framework extends to other high-stakes domains of agentic AI governance where rigorous auditability is required.

Keywords: Agentic AI governance; Verifiability Gap; delegated authority; reproducibility; explainability; FinTech.

1. Introduction

A bank that declines a loan has to be able to say why. If the applicant complains, a regulator asks, or a court becomes involved, someone must produce the reasons. If pressed, the bank must also show that the same file would be decided the same way again. That obligation is old, it is written into law, and until recently it was met by pointing at a person or at a model that could be re-run.

The systems now taking those decisions may do neither reliably. Software built on a large language model (LLM), a system trained to continue text and thereby to answer, now interprets a

goal, breaks it into steps, calls other systems, and acts with little human intervention. It has moved from pilots into production in finance. In agentic AI, the language model supplies the reasoning and language core, while tools, memory, and orchestration mechanisms enable the system to take action. These systems allocate capital, adjudicate credit, and negotiate settlements at institutional scale. Whether they can do the work is no longer the interesting question.

The interesting question is what happens afterwards. An agent may have denied a loan or filled an order. Can a supervisor, an auditor, or a court then establish *why* it acted? Can that verifier also establish *whether* the agent would act the same way again? Both answers are needed, and this paper argues that neither is currently guaranteed.

Agentic AI governance presents a new challenge that remains underexamined in IS and FinTech research. How can organizations ensure that an agent's decisions are reproducible and robust? How can they ground those decisions in an auditable reasoning process? This is not the familiar "black-box" complaint in new packaging. The black-box problem concerns opacity, and its conventional remedy is empirical validation. If a model cannot be explained, it can at least be tested: run it repeatedly, vary the inputs, and characterize its behavior. Financial institutions have validated opaque models in this way for decades, and existing supervisory practices largely assume that such testing remains possible. What is new is that this empirical remedy may no longer work. Several factors may change the result of a repeated agentic decision. The underlying model may change on the vendor's schedule, while the tools may return different information. Previously available execution controls may also be modified or withdrawn. The governance problem is therefore not opacity alone. It is the loss of reproducibility as the principal safeguard against opacity. Agentic systems can be both difficult to explain and difficult to reproduce. This conjunction and its consequences for organizational control and accountability is the central problem examined in this paper.

LLM-based and agentic AI systems are becoming central to the architecture and operation of contemporary information systems. Prior IS research has established important foundations for understanding AI management, delegation to agentic artifacts, explainability, and human oversight [Baird and Maruping, 2021, Berente et al., 2021, Fügener et al., 2021]. A longer tradition in organization and IS research examines how firms retain control over work they do not directly perform, developing formal and informal control modes and showing how control portfolios are configured and rebalanced over time [Ouchi, 1980, Kirsch, 1997, Cardinal et al., 2004, Wiener et al., 2016]. More recent work extends these questions to algorithmic and learning systems that participate in the work itself [Faraj et al., 2018, Kellogg et al., 2020, Murray et al., 2021, Teodorescu et al., 2021]. Yet this literature has only begun to theorize the distinctive governance problems posed by LLM-based agentic systems. These systems interpret goals, coordinate tools and memory, and take consequential actions with limited human intervention. We argue that much of this literature presupposes a sufficiently stable technological object that can be inspected, tested, and rerun. Agentic AI weakens this assumption because the operative decision system is distributed across models, tools, memory, orchestration mechanisms, and provider-controlled infrastructure that may change after deployment [Bommasani et al., 2021]. We therefore shift the unit of analysis from model transparency to system verifiability. We ask when delegated agentic authority exceeds an organization's capacity to explain and reproduce the resulting actions.

The consequences of weak governance for agentic AI are immediate. A reviewer who cannot follow the reasoning, question its assumptions, or reproduce the decision is not exercising meaningful oversight. The sign-off becomes a formality, and the firm assumes a liability where it believed it had established a control. An agentic decision may still be challenged after the fact. Yet if it cannot be reconstructed, corrected, or reversed within the operating system, that challenge has little practical force. Repeated failures of this kind can ultimately erode trust in the entire agentic system. The stakes are especially high in finance. Agentic AI increasingly supports capital allocation and credit

decisions, where a single action can carry both a monetary cost and a legal consequence [Gu et al., 2020, Fuster et al., 2022]. Machine-learning methods are already established in consumer credit and FinTech lending, where their distributional and regulatory consequences are actively debated [Khandani et al., 2010, Berg et al., 2020, Goldstein et al., 2019, Bartlett et al., 2022]. In this setting, explainability, reproducibility, and effective human control are not optional design choices. They are externally imposed governance obligations. Section 2 explains why finance is a critical setting for this problem.

Contributions. This paper makes three contributions to research on the governance of autonomous information systems.

First, it theorizes verifiability as an authority–evidence relation. The Verifiability Gap is not a general property of a model. It is the positive shortfall for a specific decision. On one side is the verification demanded by the authority exercised. On the other is the faithful explanation and material reproduction available to a named verifier under a given evidentiary standard and audit lag. This moves the object of governance from the model to the complete decision system and the evidence it retains.

Second, it develops a multilevel governance theory of evidence-contingent delegation. Seven propositions explain how the same authority–evidence mismatch shapes governance at three levels. Within firms, it constrains defensible delegation and control choices. Over time, it weakens historical verification after material change. Across networks, it propagates through interorganizational handoffs and shared providers. The resulting principle is that autonomous authority remains defensible only while retained verification capacity remains commensurate with the authority exercised.

Third, it makes the central reproducibility mechanisms empirically observable. We separate four targets: current outcome reproducibility, historical outcome reproducibility, material process reproducibility, and exact trace reproducibility. We use verdict differentiation to identify output collapse. Three controlled studies identify distinct mechanisms. Provider change can alter both the historical decision object and the available replay controls. Orchestration can change the operative decision rule and decouple terminal verdicts from recorded traces. Perfect current replay can also coexist with failure to recover a historical credit action. These studies identify mechanisms rather than estimate their prevalence across financial institutions.

Methodologically, Studies 1 and 2 use thirty-two fixed, constructed financial cases to isolate release, control-surface, and orchestration effects. Study 3 uses preserved versions of a transparent credit model to separate current stability, historical fidelity, and explanation faithfulness at the decision level. Figure 1 previews the argument and evidence that follow.

2. Agentic AI in FinTech

Agentic AI is shifting the role of artificial intelligence in finance from generating predictions, recommendations, and content to performing coordinated financial work. Over the coming decade, it is positioned to become a core component of the financial sector's digital intelligence workforce. This shift is already visible. JPMorgan Chase reports a wealth-management platform that includes twenty-five specialized AI agents. Visa and Mastercard have introduced infrastructure for agent-initiated commerce and payments, while FINRA identifies AI agents as an emerging focus in financial services [JPMorganChase, 2026, Visa, 2025, Mastercard, 2025, Financial Industry Regulatory Authority, 2026]. Together, these developments indicate that agentic systems are moving into consequential financial workflows.

Three distinctions are foundational to our argument. The first concerns what separates an agentic system from a conventional LLM application. The second concerns why trustworthy agentic

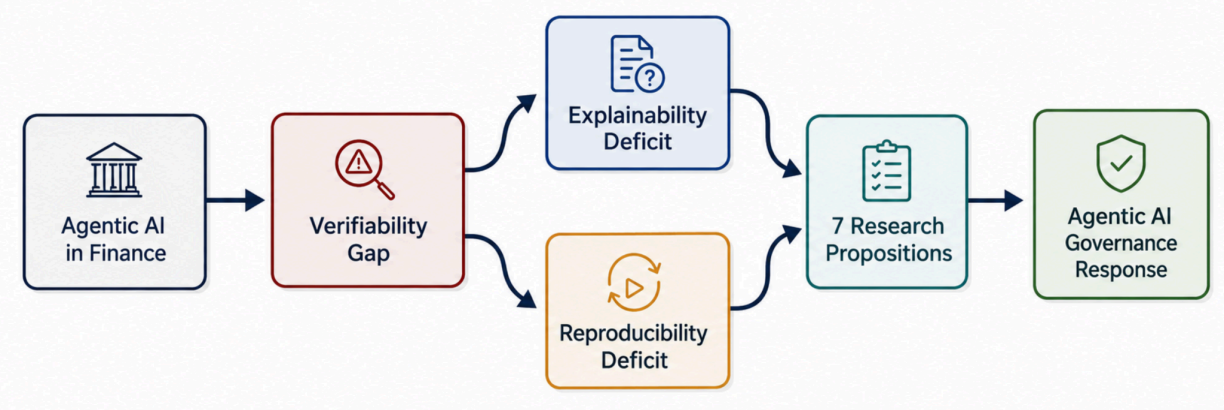


Figure 1: Roadmap of the paper's argument: from the phenomenon (agentic AI that acts in finance), to the construct (the Verifiability Gap) and its two deficits, to a falsifiable research agenda and an evidence-contingent governance response.

AI requires the responsible institution to verify how consequential actions were produced. The third concerns where the Verifiability Gap becomes most binding as decision authority shifts from people to agentic systems.

2.1 From LLMs to Agentic AI

Conventional LLM applications primarily generate content in response to user prompts. Agentic AI embeds this generative and reasoning capacity in a broader execution system that includes memory, tools, data access, execution permissions, and orchestration mechanisms. It can therefore pursue a goal through multiple decisions and actions rather than merely return an answer.

We define an *agentic AI system* as one composed of one or more LLM-based agents. These agents can (a) interpret high-level goals, (b) decompose them into executable sub-tasks, and (c) invoke external tools, models, and data. They can also (d) maintain state across steps, (e) act autonomously or semi-autonomously within delegated authority, and (f) revise their plans in response to intermediate results [Baird and Maruping, 2021, Kumar et al., 2026, Wang et al., 2023]. Interleaving generated reasoning with tool invocation is the mechanism that turns a language model into an actor rather than a responder [Wei et al., 2022, Yao et al., 2022].

The enterprise frontier is multi-agent orchestration. Specialized agents may perform quantitative analysis, compliance review, information retrieval, customer interaction, and execution. They may interact through sequential, directed-acyclic, or cyclic communication structures [Rai et al., 2019, Xiao et al., 2024].

The argument that follows turns on a small number of operating details of these systems. Table 1 states each one in plain terms, and notes what it governs. Three of them are the recurring subject of this paper: what the institution can *keep*, what it can *re-run*, and what it can only *observe*.

This architecture changes the object of governance. A consequential financial action may emerge from a distributed sequence of model states, prompts, retrieved records, memory updates, tool responses, inter-agent messages, handoffs, and execution steps. A single agent's trace may therefore be insufficient to explain the joint outcome. Financial work becomes distributed across the agentic system, while organizational and legal accountability remains with the institution that authorized it. This separation creates the architectural conditions under which the Verifiability Gap can widen, and it is a recognized source of harm as systems are granted progressively more agency [Chan et al.,

Table 1: Operating vocabulary used throughout the paper, and why each term matters for governance.

Term	**What it is**	**Why it matters here**
Large language model (LLM)	Software that writes text by repeatedly predicting the next small fragment of text, given everything before it.	The reasoning and language core; it decides nothing on its own.
Prompt	The text handed to the model on one call: the instructions plus the case at hand.	Part of the evidence a verifier needs; if it is not kept, the call cannot be reconstructed.
Sampling	Choosing the next fragment at random from the model's ranked candidates rather than always taking the top one.	The reason two runs of one input can differ at all.
Decoding settings	The provider-exposed dials governing that choice — here, *temperature* and *random seed*.	The institution's only levers on reproducibility, and it owns neither.
Tool call	A request the model emits for the surrounding system to act — fetch a record, price an instrument — and return the result.	Brings outside state into the decision; that state must be kept to replay it.
Handoff	The transfer of one agent's output to the next agent as its input.	Where evidence is compressed, and where the paper locates evidentiary loss.
Execution trace	The ordered record of one decision: every agent output, tool call, handoff, and the final action.	The object a verifier reconstructs. Its absence is the Verifiability Gap.
Release (model version)	The dated build of the provider's model actually served when the decision was made.	Changes on the provider's schedule, not the institution's.
Executable	The specific fitted model, code, and settings that produced a past decision — not whatever is deployed today.	Historical verification is indexed to this, not to the current system.
Replay	Re-running a preserved decision on preserved inputs to see whether the same action results.	The empirical route to verification when explanation is unavailable.
Modal verdict	The action returned most often across repeated executions of the same case.	The summary this paper's reproducibility measures compare against.

2023].

2.2 The Verifiability Mandate

Trustworthy-AI research in finance commonly emphasizes explainability, fairness, robustness, security, privacy, and accountability. These properties remain essential. Yet for the institution held responsible for an agentic action, the central governance question is evidentiary: can it establish that these properties held in the particular decision under review? This requirement is reinforced by existing expectations concerning model-risk management, documentation, oversight, and accountability [Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011, Financial Industry Regulatory Authority, 2024, European Parliament and Council, 2024]. A parallel literature asks what it takes to hold algorithmic systems accountable in practice, and converges on retained evidence rather than disclosure alone: accountability requires records that a reviewer can actually interrogate [Diakopoulos, 2016, Wieringa, 2020], internal audit procedures that run before and after deployment [Raji et al., 2020], and decision systems designed from

the outset to be reviewable [Cobbe et al., 2021].

Explainability and reproducibility provide two foundational conditions for answering this question. Explainability concerns whether the institution can establish *why* an action was taken. Reproducibility concerns whether it can reconstruct *how* the action was produced and determine whether the relevant method and outcome can be reproduced under documented conditions [Gregor and Benbasat, 1999, Rudin, 2019, Pineau et al., 2021, Atıl et al., 2025].

For agentic AI, both requirements extend beyond the final LLM output. They require a sufficient record of the operative model version and state, prompts, retrieved information, memory, tool calls, execution parameters, inter-agent communications, intermediate decisions, and resulting actions. A repeated outcome without a reconstructable decision process provides only partial evidence. Likewise, a plausible explanation generated after the event is inadequate if it cannot be connected to the process that produced the action.

Consider an agentic trading system that submits an order. Aggregate performance or risk measures do not explain why that particular trade occurred. The firm may need to reconstruct the market data, risk constraints, prompts, analytical steps, agent handoffs, tool calls, and routing decisions that produced the execution. Confirming only that the order occurred does not constitute meaningful oversight.

Verifiability is therefore not a substitute for trustworthiness. It is a foundational evidentiary condition under which trustworthiness can be demonstrated, audited, and contested. A system may appear fair, robust, and secure in aggregate. Yet if its consequential actions cannot be explained or reproduced, the responsible institution cannot establish whether those properties held in the specific case being challenged. Without that evidence, human review risks becoming ex post approval rather than meaningful organizational control.

2.3 Where the Verifiability Gap Binds: Domain and Delegated Autonomy

The Verifiability Gap does not arise uniformly across FinTech applications. Its importance depends not only on the financial domain, but also on how much decision authority the institution delegates to the agentic system. We therefore organize representative applications along two dimensions: *application domain* and *delegated autonomy*. This is an analytical map rather than an exhaustive catalog.

We distinguish four levels of delegated autonomy. At the *assist* level, the agent retrieves, organizes, or summarizes information for a human decision maker. At the *recommend* level, it proposes a course of action, but a human retains formal authority over the decision. At the *act-with-approval* level, the agent executes an action after a human checkpoint. At the *act-autonomously* level, it executes within a bounded mandate without case-by-case human authorization. Autonomy in this framework therefore refers to delegated decision authority, not to the absence of organizational constraints.

Verifiability pressure generally rises as authority moves from the human to the agentic system. It is lower, but not absent, at the assist and recommend levels, particularly when human reviewers defer to agent recommendations [Logg et al., 2019]. It becomes most consequential at the two act levels, where the institution remains accountable for an action that it did not directly author.

Table 2 maps representative FinTech applications across this domain-by-autonomy space.

Three implications follow from this map. First, the same financial domain can contain materially different governance conditions. An agent that summarizes a credit file and an agent that denies the application may use similar models and data, but they exercise different levels of institutional authority. The Verifiability Gap is therefore not a property of the application domain alone.

Second, the use of agents does not by itself determine the size of the gap. The gap depends on the

Table 2: Illustrative agentic AI applications in FinTech by domain and delegated autonomy. Verifiability pressure generally rises from left to right as more decision authority is delegated to the agentic system. (AML: anti-money laundering; A2A: agent-to-agent.)

Increasing delegated authority and verifiability pressure $\longrightarrow$				
Domain	**Assist**	**Recommend**	**Act with approval**	**Act autonomously**
Wealth management	Surface portfolio, tax, and market information	Propose allocation or rebalancing	Execute rebalancing after sign-off	Conduct bounded rebalancing and tax-loss harvesting
Compliance / AML	Summarize alerts and retrieve evidence	Recommend alert disposition	Escalate or close cases after approval	Continuously monitor and take bounded case actions
Credit origination	Retrieve data and draft credit memoranda	Recommend approval, pricing, or limits	Issue approved terms or decisions	Approve, deny, or price credit within mandate
Trading	Retrieve research and generate signals	Propose trades or strategies	Submit approved orders	Execute and adapt strategies within risk limits
A2A settlement	Match records and identify breaks	Propose settlement terms	Submit an approved settlement	Negotiate and settle directly with other agents

relationship between delegated authority and verifiable control. Durable records, stable execution controls, and reconstructable decision paths can make a highly autonomous system verifiable. By contrast, a less autonomous system may be difficult to verify when its recommendations are routinely accepted without meaningful review.

Third, the economic promise of agentic AI partly lies in moving financial work rightward across the table. Multi-agent trading systems, for example, are designed not only to deliberate but also to execute resulting trades [Xiao et al., 2024]. Yet the strongest expectations concerning documentation, model-risk management, auditability, and institutional accountability apply precisely where the agentic system exercises the greatest authority [Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011, Financial Industry Regulatory Authority, 2024]. The benefits of delegated execution and the need for verifiable control therefore increase together.

The remainder of the paper focuses primarily on settings in which agents perform or materially determine consequential financial actions. These settings occupy the act-with-approval and act-autonomously portions of the map. Here, verifiability becomes a binding organizational control problem rather than merely a desirable system property.

3. Research Approach: FinTech Agentic AI Verifiability Gap

Our core question concerns the boundary of defensible delegation. When does the authority delegated to a FinTech agentic AI system exceed an organization's capacity to explain and reproduce its actions? Our unit of analysis is the complete agentic decision system: models, tools, memory, orchestration, and controls, not just the isolated underlying model. The research follows three linked stages:

1. *Phenomenon Specification:* Defining the analytical boundary within FinTech.
2. *Theory Construction:* Theorizing the Verifiability Gap as the joint failure of two partially compensating verification routes: explainability and reproducibility.
3. *Controlled Mechanism Tests:* Observing the theory's reproducibility implications.

This design aims to identify mechanisms, rather than provide an exhaustive review or estimate population-level prevalence.

3.1 Phenomenon Specification

We focus on LLM-based agentic systems that execute or drive significant financial actions, specifically at the act-with-approval or act-autonomously levels. At these stages, institutions remain accountable for decisions generated by distributed technical systems. Literature and regulatory sources highlight the shift from content generation to coordinated action. Our goal is to theorize the mechanisms that affect verifiability, not to catalog all agentic AI applications.

3.2 Theory Construction

We build our theory by combining two concepts that are typically examined separately:

The explainability deficit. A fluent, plausible reasoning record may not faithfully represent the actual decision-making process [Gregor and Benbasat, 1999, Jacovi and Goldberg, 2020, Lanham et al., 2023, Turpin et al., 2023]. Interpretability itself remains contested in both definition and evaluation [Lipton, 2018, Doshi-Velez and Kim, 2017], and widely used post hoc attribution methods can fail basic faithfulness checks [Ribeiro et al., 2016, Lundberg and Lee, 2017, Adebayo et al., 2018].

The reproducibility deficit. Re-executing a task under similar conditions may not yield the same outcome or decision process [Pineau et al., 2021, Kapoor and Narayanan, 2023, Atıl et al., 2025, Han, 2025]. This concern is long-standing in computational and machine-learning research, where reported results have proven difficult to independently reproduce and where uncontrolled sources of variance routinely exceed the effects under study [Peng, 2011, Gundersen and Kjensmo, 2018, Raff, 2019, Bouthillier et al., 2021].

The Verifiability Gap occurs when delegated authority expands while both verification routes fail: the organization can neither establish why an action was taken nor reconstruct how it was produced. Existing constructs such as the responsibility gap, inscrutability, and decision provenance emphasize only one part of the problem. The Verifiability Gap instead theorizes the joint failure of explainability and reproducibility at the agentic-system level. Table 3 summarizes this distinction.

Table 3: Positioning the Verifiability Gap against adjacent constructs. Existing constructs emphasize one verification route or treat reproducibility as given; the Verifiability Gap theorizes their joint loss at the agentic-system level.

Construct	**Explainability route**	**Reproducibility route**	**Agentic-system focus**
Responsibility gap [Matthias, 2004]	Implicit	No	No
Inscrutability [Berente et al., 2021]	Yes	No	Partial
Envelopment [Asatiani et al., 2021]	Yes	No	No
Provenance [Singh et al., 2019]	Partial	Assumed	No
Contestability [Alfrink et al., 2023]	Yes	No	No
Verifiability Gap (this paper)	**Yes**	**Yes**	**Yes**

This synthesis yields seven propositions organized across three governance levels. P1–P3 address defensible delegation and control choices within the firm. P4–P5 address historical replay and evidentiary validity after material change. P6–P7 address end-to-end and common-mode exposure across organizational networks. The controlled studies examine central mechanisms underlying P4 and P5 and a system-internal analogue of the serial dependence in P6. The remaining organizational and network consequences are specified as falsifiable propositions for subsequent field research.

3.3 Controlled Mechanism Tests

The empirical component comprises three linked but analytically distinct mechanism tests. Studies 1 and 2 share thirty-two fixed FinTech cases covering credit, trading, anti-money-laundering review, and portfolio rebalancing. Study 1 varies dated provider releases and the execution controls available for replay. Study 2 holds the locally served model, cases, and observed execution settings fixed while changing the agentic architecture from one to fifty agents. Study 3 freezes two versions of a transparent credit model and asks whether current stability, historical decision recovery, and explanation faithfulness can diverge within one decision episode.

The studies operationalize central reproducibility mechanisms of the Verifiability Gap rather than the full gap itself. They distinguish four targets: four distinct targets, defined in Section 7.1. Verdict differentiation is tracked separately as a diagnostic of output collapse. This decomposition prevents three distinct phenomena from being conflated. A current verdict may repeat even when the historical action is no longer recoverable. A terminal action may recur while the recorded execution path changes. Apparent stability may also improve because the system has stopped distinguishing among cases.

The design is deliberately bounded for mechanism identification rather than population estimation. Figure 2 summarizes the research logic.

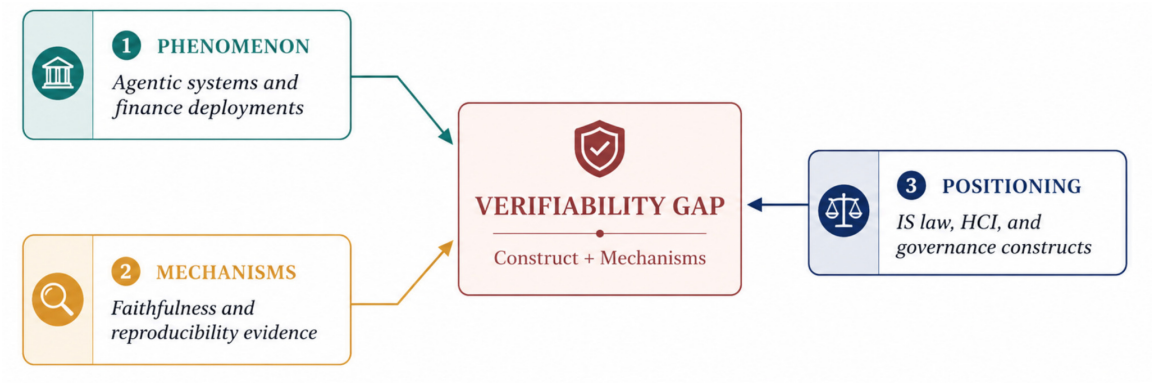


Figure 2: Research design. Phenomenon specification defines the analytical boundary in FinTech; theory construction joins explainability and reproducibility in a seven-proposition architecture; and three controlled studies examine central mechanisms through which retained verification capacity can fall behind delegated authority.

4. The Verifiability Gap: Definition and Generative Mechanisms

Agentic systems shift financial AI from merely recommending actions to autonomously executing them. This shift raises a critical governance question: can an institution later establish *how* that delegated authority was exercised?

Prior research often treats opacity as the central governance problem. In agentic AI, however, the issue is not simply whether the system is transparent. The institution remains responsible for actions taken under the authority it delegates to the system. It must therefore be able to explain why those actions were taken and reproduce the process that produced them. The Verifiability Gap opens when delegated authority exceeds this verification capacity. This mismatch remains underexamined in IS and FinTech research on agentic AI.

Verifiability Gap. We define the Verifiability Gap as a shortfall in retained verification capacity.

The gap arises when delegated agentic authority demands more verification than the organization retains. That capacity must support both a faithful explanation of the resulting action and a material reproduction of the process that produced it.

An Example. Consider an agentic system that evaluates loan applications. When the agent merely recommends a denial for a human officer's review, the delegated authority is low, and basic logging may suffice.

However, if the institution authorizes the agent to execute the denial autonomously, the verification requirement sharply increases. A Verifiability Gap emerges if the bank cannot recover the exact model version, tool outputs, and execution path necessary to faithfully explain and reproduce that denial. The loan outcome itself may be identical, but the gap widens because the delegated authority expanded without a commensurate increase in preserved verification capacity.

Let d denote a decision episode and $q = (v, \sigma, \tau)$ its audit context, comprising the verifier (v), evidentiary standard (σ), and elapsed time since execution (τ). Let A_d denote the authority exercised, $\rho_\sigma(A_d)$ the verification required for that authority, and V_{dq} the verification capacity retained by the system. The gap is:

$$\boxed{G_{dq} = \Big[\underbrace{\rho_\sigma(A_d)}_{\text{verification required}} - \underbrace{V_{dq}}_{\text{verification available}} \Big]_+} \tag{1}$$

where $[x]_+ = \max\{0, x\}$. Both terms are assessed on a common bounded scale defined by the evidentiary standard.

Equation 1 distinguishes two concepts that prior frameworks often conflate: the verification *required* by delegated authority and the verification capacity *retained* by the system. Delegated authority is not itself a deficit. It sets the evidentiary burden that the institution must meet after an action is taken. A Verifiability Gap becomes binding when retained capacity falls below this requirement. In our framework, this occurs when both routes to verification: faithful explanation and material reproduction are impaired. If either route remains sufficient, the institution faces a localized deficit rather than the full Verifiability Gap.

The construct therefore captures the institution's shortfall in substantiating how delegated authority was exercised, independent of whether the resulting decision happened to be accurate, fair, or profitable. As Figure 3 illustrates, the gap is a joint failure. The value of G_{dq} formalizes the resulting organizational exposure. When $G_{dq} = 0$, retained evidence meets the verification requirement created by the delegated authority. When $G_{dq} > 0$, its value represents the magnitude of the verifiability shortfall and the associated governance exposure.

4.1 What the Verifiability Gap Measures: Specific Financial Decisions

The formula does not measure an agentic system's general verifiability. It measures the verification shortfall produced by the complete FinTech agentic system in one specific financial decision. In the examples above, the relevant unit is therefore not the lending, trading, or fraud-management system in general, but a particular loan denial, executed trade, or fraud disposition. Each decision may involve different data, tool calls, system states, delegated authority, and retained evidence.

Accordingly, d denotes a specific financial decision episode. The episode includes the case presented to the system, the decision process executed, the resulting financial action, and the evidence retained for subsequent verification. The *complete FinTech agentic system* includes its model, data sources, tools, memory, orchestration, and execution controls. This system produces the decision episode. It therefore determines both the authority exercised, A_d, and the verification capacity later

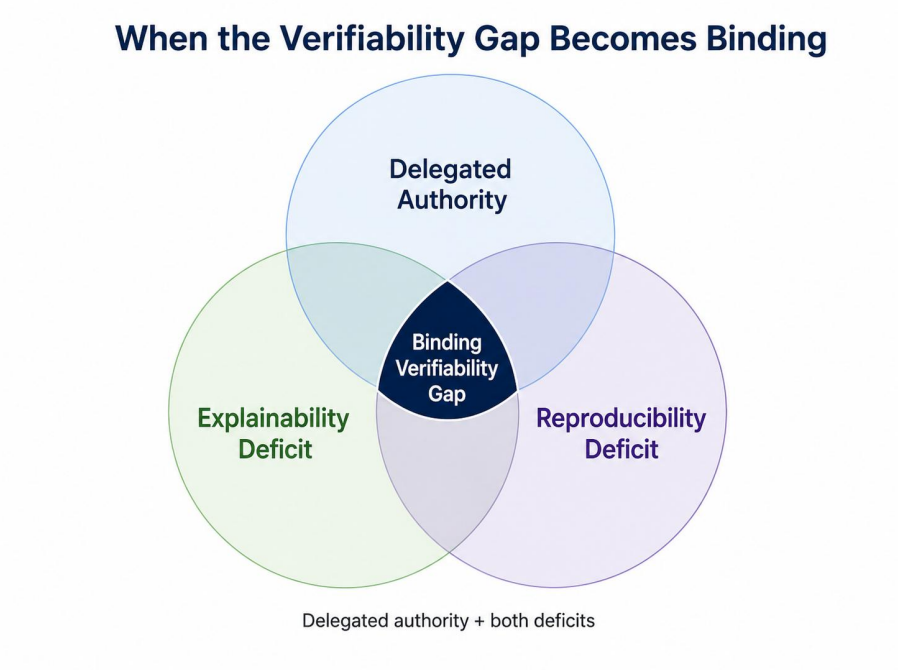


Figure 3: Conditions under which the Verifiability Gap becomes binding. The gap is most consequential when delegated agentic authority coincides with deficits in both explainability and reproducibility.

available, V_{dq}. Thus, the complete system is the source of the Verifiability Gap, while the specific financial decision is the unit for which the gap is calculated.

4.2 Verification Capacity in the Verifiability Gap: Explainability and Reproducibility

Consider again the loan denial described above. The institution may possess two distinct forms of evidence for verifying that decision. The first route is *explainability*, represented by E_{dq}. It requires evidence that connects the denial to the applicant data, tool outputs, and lending rules actually used. A rationale generated after the decision is insufficient unless it faithfully corresponds to the executed process. Counterfactual accounts of the form "the outcome would have differed had this factor been different" have been proposed precisely for automated decisions of this kind [Wachter et al., 2017], and they presuppose a decision function that can still be interrogated.

The second route is *reproducibility*, represented by R_{dq}. It requires evidence that reconstructs the materially relevant decision process. That evidence includes the recorded application data, credit and income-verification outputs, policy and model versions, tool calls, and execution states. Producing the same denial through different data, rules, or controls does not reproduce how the delegated authority was originally exercised.

These forms of evidence are distinct. A bank may preserve a faithful explanation for the denial while lacking the data or system states required to reconstruct it. Conversely, it may rerun the recorded process and reproduce the denial while remaining unable to establish why particular factors determined the result. Verification capacity is therefore:

$$V_{dq} = V_\sigma(E_{dq}, R_{dq}),$$

where the evidentiary standard σ determines whether explainability, reproducibility, or both are

required to verify decision d.

4.3 Why Serial FinTech Agent Architectures Widen the Verifiability Gap

Serial dependence is causal dependence. Serial dependence is a central architectural mechanism that can widen the Verifiability Gap in FinTech. For a financial decision episode d, a *serial dependency* exists when a state produced by one agent becomes a causally material input or constraint for a downstream agent. A transferred state is causally material when, under a relevant counterfactual, changing that state could change a downstream assessment, control determination, or final financial action. Seriality therefore concerns causal dependence, not merely temporal order or agent count.

Serial dependence is broadly relevant because many consequential FinTech workflows decompose financial decisions into specialized stages, including data acquisition, verification, risk analysis, policy review, and execution. Recent agentic designs similarly organize financial tasks around specialized roles and hierarchical handoffs [Xiao et al., 2024, Kumar et al., 2026]. The mechanism is not limited to a purely linear chain. It also applies to hierarchical and mixed serial–parallel architectures whenever the executed financial action depends on at least one causally material serial path.

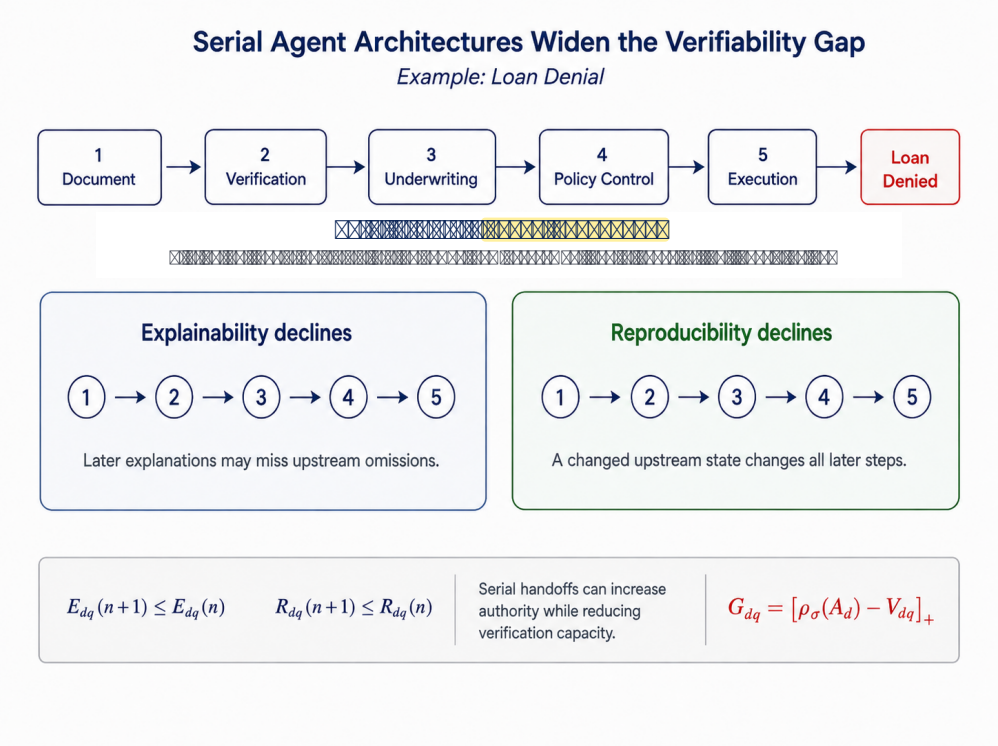


Figure 4: Serial dependence in a FinTech loan-denial workflow. Causally dependent handoffs can weaken end-to-end explainability and reproducibility, reducing verification capacity while broader delegated authority increases verification demand.

Consider a loan-denial episode. A verification agent omits a source of variable income when summarizing an income-verification response. The downstream underwriting agent consequently calculates a debt-to-income ratio above the lending threshold, a policy-control agent accepts that assessment, and an execution agent denies the loan. The final explanation may faithfully report that the application was denied because the ratio exceeded the threshold. Yet it fails to reveal that

the ratio depended on an upstream omission. The explanation is therefore locally faithful to the terminal decision while remaining globally incomplete as an account of the institutional action. If the original tool response and transferred state were not retained, the institution may also be unable to reconstruct the process that produced the denial. An insufficiently captured serial handoff can thus weaken both routes to verification simultaneously. Figure 4 summarizes this mechanism.

Serial decay of explainability. End-to-end explainability requires a faithful evidentiary link from the final financial action back through every causally material transformation. A terminal agent's rationale, standing alone, accounts only for the mapping from the state it received to the action it selected. As information moves through serial handoffs, underlying evidence may be compressed into summaries, interpretations, or intermediate judgments. Upstream omissions can therefore become embedded premises for downstream agents. Unless these transformations are preserved and linked to the final action, an explanation can remain locally plausible while failing to account for the end-to-end decision process.

Serial decay of reproducibility. End-to-end reproducibility requires the materially relevant states to be reconstructed in the sequence in which they shaped the financial action. Every downstream agent conditions its output on a state produced upstream. If an early retrieval, transformation, summary, or tool response diverges, subsequent agents operate on a different context. A fixed random seed cannot repair this dependence: applying the same seed to a different conditioning state does not recreate the original decision process.

To state this mechanism precisely, let d_n and d_{n+1} denote counterfactually comparable decision episodes that differ only in the addition of one causally material serial transition. We hold the financial case, delegated authority, audit context, and verification controls constant. We also assume that the added transition provides no compensating verification artifact:

$$E_{d_{n+1}q} \leq E_{d_n q}, \qquad R_{d_{n+1}q} \leq R_{d_n q}. \tag{2}$$

The inequality is strict for explainability when the added transition loses or obscures causally relevant evidence. It is strict for reproducibility when the transferred state cannot be reconstructed under the recorded conditions. Equality is route-specific. The added transition must preserve either the evidence needed for explanation or the state needed for reproduction. It must do so with enough fidelity to satisfy the applicable evidentiary standard.

The theoretical driver is therefore dependence, not agent count. Parallel agents whose outputs are independently retained do not incur this specific cumulative dependence across branches. However, when a downstream agent summarizes, filters, approves, or executes their combined outputs, the aggregation step reintroduces serial dependence.

The Double Effect. The theoretical significance of seriality lies in its potential *double effect*. By coordinating specialized agents across data collection, risk assessment, control review, and execution, a serial architecture can enable institutions to delegate broader and more consequential financial authority, increasing $\rho_\sigma(A_d)$. At the same time, every causally material handoff adds another state whose evidentiary content must be preserved for explainability and whose execution conditions must be reconstructed for reproducibility. Failures of preservation can therefore reduce E_{dq} and R_{dq}, and consequently lower V_{dq}. When broader delegation and weaker verification capacity occur together, serial architectures widen the Verifiability Gap from both sides: they increase the verification demanded while reducing the verification capacity retained.

4.4 Governance Requires a Reversible Chain

The double effect can be stated exactly. Doing so converts the verbal mechanism into a constraint on architecture, and it supplies the common scale that Equation 1 presumes. Proofs are in Appendix S2.

Index the model calls in a topological order of the communication graph, write S_k for the outputs of layer k, T for the complete trace, and $Y = g(T)$ for the financial action. The governance design imposes one restriction that we use directly: *only the first layer observes the original case*, and every downstream agent observes only the records transmitted along designated edges. Writing C for the case, the joint law therefore factorizes as a Markov chain,

$$C \longrightarrow S_1 \longrightarrow S_2 \longrightarrow \cdots \longrightarrow S_d \longrightarrow Y. \tag{3}$$

This is not an assumption adopted for tractability. It is the information-access rule the architecture enforces, and it is what makes an agentic pipeline auditable at all.

Auditing runs in the opposite direction from execution. A verifier is handed an action and asked to establish how it was produced. That is an inverse problem on (3), and the property it demands has a name.

Definition 1 (σ-reversibility)**.** Let $\equiv_\sigma$ denote material equivalence under the evidentiary standard σ, with classes $[\,\cdot\,]_\sigma$. Stage k is *σ-reversible given retained record E_k* if $(S_k, E_k) \mapsto [S_{k-1}]_\sigma$ is well defined almost surely. A configuration is σ-reversible if every causally material stage is, so that the verifier can reconstruct $[T]_\sigma$ from Y and the retained bundle E.

Remark 1. We do not mean time-reversibility in the detailed-balance sense. The relevant question is whether the earlier state can be recovered from what was retained.

The next result uses one standard quantity. Write $I(C;Y)$ for the *mutual information* between the case and the action. Informally, it is how much the executed action tells you about which case was being decided. It is zero when the action is the same whatever the case. It is largest when each case draws its own action.

The result rests on one classical fact, the *data-processing inequality.* Passing a signal through a further stage of processing cannot create information the stage never received. The coefficient η_k below records how much of that information a given layer lets through. At $\eta_k = 1$ the layer loses none.

Theorem 1 (Serial contraction of discrimination)**.** *Under* (3)*, $I(C;Y) \le I(C;S_d) \le \cdots \le I(C;S_1) \le H(C)$. If η_k denotes the strong data-processing coefficient of the layer-k channel, then $I(C;Y) \le \big(\prod_{k\ge 2}\eta_k\big) I(C;S_1)$, so discrimination decays geometrically in depth whenever $\eta_k \le \bar\eta < 1$.*

In plain terms: what the final action can reveal about which case was decided never grows as the pipeline deepens, and shrinks geometrically once each layer loses anything.

Corollary 1. *No downstream agent, aggregation rule, critic, or terminal decider can raise $I(C;Y)$ above $I(C;S_1)$. Case information not captured by the first layer is unrecoverable by any later stage, however capable.*

Theorem 2 (Reversibility is equality in the data-processing inequality)**.** *With no retained record, the following are equivalent: (i) stage k is σ-reversible; (ii) $I(C;S_k) = I(C;S_{k-1})$, i.e. $\eta_k = 1$; (iii) S_k is a* sufficient statistic *for C relative to S_{k-1}, carrying everything the upstream state carried about the case; (iv) the stage map is injective on the support of S_{k-1} modulo $\equiv_\sigma$.*

In plain terms: a stage is auditable exactly when it has thrown nothing away. Auditability and information preservation are not two related properties but one. The next result says what that costs.

Theorem 3 (Specialization and reversibility are incompatible)**.** *If stage k performs summarization in the ordinary sense—deterministic and not injective modulo $\equiv_\sigma$—then $\eta_k < 1$ strictly, the stage is not σ-reversible with a trivial record, and there exist materially distinct upstream states the verifier cannot separate from the downstream record. Conversely, a stage that is σ-reversible with a trivial record performs no material compression; it merely relabels.*

In plain terms: a stage either summarizes or stays auditable. It cannot do both.

This is the double effect as a dichotomy. The operation that makes specialized delegation worthwhile—compressing an upstream state into an actionable summary—is precisely the operation that destroys reconstructability. An architecture cannot obtain the first without incurring the second. The choice available to an institution is therefore not whether to lose reversibility, but whether to pay to restore it.

One further symbol is needed. Write $H(X)$ for the *entropy* of a quantity X: the average number of bits required to record which value X took. It is zero when X never varies. It grows with the number of distinct values X takes and how evenly it takes them. $H(X \mid Z)$ is the same count once Z is already known. So $H([T]_\sigma \mid Y)$ below counts the bits still needed to pin down the material process after the action is known.

Theorem 4 (Retention lower bound)**.** *A configuration is σ-reversible given E only if*

$$H(E) \;\geq\; H\big([T]_\sigma \,\big|\, Y\big), \tag{4}$$

and the bound is achieved by retaining, at each stage, an index of the material equivalence class of the upstream state within the pre-image of the downstream state. The evidence bundle must be at least as large as the information the pipeline destroys.

In plain terms: to keep a decision reconstructable, the institution must store at least as many bits as its pipeline threw away.

Corollary 2 (The Verifiability Gap in bits)**.** *Instantiating the verification required by the authority exercised as the reconstruction requirement $H([T]_\sigma \mid Y)$, and the verification capacity retained as $H(E)$,*

$$G_{dq} \;=\; \Big[\, H\big([T]_\sigma \mid Y\big) - H(E) \,\Big]_+ \tag{5}$$

is well defined in bits and equals zero exactly when the configuration is σ-reversible given what was kept.

Corollary 2 constructs the common bounded scale that Equation 1 presumes. The two sides become commensurable because both are measured in the same units: the information a verifier needs, and the information the institution kept.

Corollary 3 (Retention must grow with depth)**.** *Each materially compressing stage contributes a strictly positive term to* (4)*. The retention required to hold $G_{dq} = 0$ is therefore non-decreasing in the number of causally material stages and strictly increasing in the number that compress. Deeper orchestration does not merely risk auditability; it raises the evidentiary price of the same autonomy.*

5. The Verifiability Gap in FinTech AI Agents: Scope and Governance Threshold

The Verifiability Gap must therefore be governed in FinTech. Otherwise, agentic systems may exercise consequential financial authority without sufficient explainability or reproducibility. Institutions and regulators would then be unable to substantiate those decisions. We therefore specify when the gap applies and the threshold at which it emerges.

The Verifiability Gap asks whether a named verifier, such as an internal risk committee or external regulator, can obtain sufficient evidence to substantiate a delegated financial decision. It differs from three related ideas. The *responsibility gap* concerns who can be held responsible when autonomous systems contribute to harm [Matthias, 2004]; the Verifiability Gap can exist even when a decision is accurate and profitable. *Inscrutability* concerns whether a system can be understood [Berente et al., 2021]; the Verifiability Gap concerns whether its decision can be substantiated through sufficient explainability or reproducibility. *Provenance* records the origin and custody of artifacts [Singh et al., 2019]; the Verifiability Gap asks whether those artifacts are sufficient to substantiate a specific decision. Finally, while model-risk management assumes that independent validation remains possible [Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011], the Verifiability Gap identifies when retained verification capacity no longer meets the level required by delegated authority.

Explainability and reproducibility can compensate for one another. A Verifiability Gap therefore arises only when neither is sufficient to meet the verification required for the delegated decision.

The Verifiability Gap applies within four boundaries:

1. It concerns *ex post* verification of a completed financial decision, not *ex ante* model testing.
2. It concerns the ability to substantiate a decision, not whether the decision is accurate, fair, legal, or profitable.
3. It applies only when an agentic system exercises substantive delegated authority. Nominal or "rubber-stamp" human approval does not eliminate the gap, since placing a person in the loop does not by itself produce substantive review [Green and Chen, 2019].
4. It is assessed for a specific decision episode and audit context, not as an inherent property of a model.

Equation 1 defines the governance threshold:

$$G_{dq} = 0 \quad \Longleftrightarrow \quad V_{dq} \geq \rho_{\sigma}(A_d). \tag{6}$$

Conversely, a Verifiability Gap exists, $G_{dq} > 0$, when $V_{dq} < \rho_{\sigma}(A_d)$.

This threshold does not require financial institutions to minimize AI autonomy. It requires the verification capacity retained after a decision to meet or exceed the verification required by the authority exercised. The threshold identifies three direct governance levers:

- *Lower the requirement:* Narrow the financial authority that an agentic system can exercise without substantive human review.
- *Raise explainability:* Preserve artifact-grounded traces sufficient to substantiate the end-to-end decision process.
- *Raise reproducibility:* Preserve model and tool versions, financial data, configurations, intermediate states, and execution conditions required for a materially equivalent rerun.

The propositions that follow explain how delegated authority, explainability, and reproducibility change within firms, during regulatory supervision, and across inter-firm agent-to-agent (A2A) networks.

6. From the Verifiability Gap to Multilevel Governance Consequences

The Verifiability Gap is a governance exposure, not a rebranded AI risk. It emerges when the verification demanded by delegated authority exceeds an institution's retained capacity to substantiate that authority. This capacity relies on two complementary routes: explainability and reproducibility.

Delegated financial authority creates an evidentiary demand, $\rho_\sigma(A_d)$. Conversely, the operative agentic system determines the actual retained evidence. This system comprises models, tools, memory, orchestration, provider interfaces, and organizational controls. This architecture dictates whether an institution can substantiate a challenged decision.

Thus, our central claim goes beyond "agentic AI is opaque" or "LLMs are nondeterministic". An agentic system can remain highly capable, accurate, and behaviorally stable. Yet, the institution may still become entirely unable to substantiate how that system exercised its delegated authority.

Seven propositions. To address this governance exposure, we advance seven propositions tracing how this authority–evidence mismatch produces three classes of consequences: firm-level, regulatory-level, and network-level.

- *The firm level (P1–P3)* explains how the gap constrains delegated authority and control choices inside the FinTech firm.
- *The regulatory level (P4–P5)* explains why compliance that appears adequate at deployment can lose substantive force during an audit.
- *The network level (P6–P7)* explains how local verification deficits cascade into end-to-end and common-mode exposures across organizational networks.

These are not seven independent risks. They are multilevel governance consequences of the same fundamental gap.

6.1 Firm-Level Consequences

At the firm level, the Verifiability Gap has three structural consequences. It caps defensible delegation, reweights the value of governance controls, and creates a tradeoff between verifiability and efficiency.

Delegated authority is bounded by institutional defensibility. Technical capability dictates what an agentic system *can* do, but institutional defensibility dictates what it *may* be authorized to do. Let c denote maximum technical autonomy, a granted autonomy, V retained verification capacity, and κ task consequence. The maximum defensible autonomy is:

$$a^*(c, V, \kappa) = \sup \{a \leq c : \rho_\sigma(a, \kappa) \leq V\}. \tag{7}$$

Equation 7 establishes an *authority ceiling*. Beyond this point, new technical capabilities expand feasible actions but not defensible ones. This ceiling rises with greater verification capacity and falls as financial stakes increase [Baird and Maruping, 2021, Tibebu, 2026].

P1 (Firms delegate only what they can later defend). Conditional on technical capability, a firm's retained verification capacity shapes the autonomy it grants to a FinTech agentic system. Autonomy will increase with that capacity until the technical ceiling is reached. When that capacity remains fixed, granted autonomy will plateau as capability rises, and the plateau will be lower for more consequential financial tasks.

Uncertain explainability changes the relative value of controls. Explainability only aids review if the disclosed rationale is causally faithful. Let ϕ denote confidence in faithfulness. Let δ_F and δ_U denote the probabilities that explanation-based review detects and prevents a fault before execution under faithful and unfaithful explanations, respectively. Let b denote the probability that an action bound blocks the fault. Action bounds become the superior safeguard when:

$$b > \phi\delta_F + (1-\phi)\delta_U, \qquad \delta_F > \delta_U. \tag{8}$$

As confidence (ϕ) drops, action bounds become relatively more dependable than explanation-based review (Equation 8). Fluent but unfaithful explanations exacerbate this by making faulty decisions appear considered, suppressing reviewer scrutiny and driving δ_U even lower [Goddard et al., 2012, Turpin et al., 2023]. While explainability and bounds remain complements, uncertain faithfulness forces the governance burden onto hard constraints.

P2 (When explainability is unreliable, firms must constrain actions). As uncertainty about explanation faithfulness increases, action-bounding controls will outperform review based on explanation records alone. Specifically, they will produce a lower consequential fault-escape rate. The difference will be greatest when the record is fluent but unfaithful.

Verifiable autonomy has an operating cost. Controls that preserve verifiability include version pinning, environment replay, deterministic settings, and substantive human approval. These controls reduce the operational speed and flexibility that make agentic AI attractive. Let x be control intensity, m the task's adaptation requirement, $V(x;\theta)$ verification capacity, and $Y(x,m;\theta)$ operating performance. Within a fixed architecture:

$$V_x > 0, \qquad Y_x < 0, \qquad Y_{xm} < 0, \qquad \left.\frac{dY}{dV}\right|_{\theta,m} = \frac{Y_x}{V_x} < 0. \tag{9}$$

Equation 9 defines a *verifiability–efficiency frontier*. While better architectures can push this frontier outward, firms cannot treat verifiability as costless. This operational tax is greater for tasks requiring frequent adaptation.

P3 (Making agents verifiable costs speed and flexibility). Within a fixed technical architecture, stronger controls that preserve explainability and reproducibility impose an operating cost. They reduce the speed, flexibility, or scope of autonomous action. The operating penalty will be greater for tasks requiring frequent adaptation.

This tradeoff compounds dynamically. Substantive human review is a core component of verification capacity. If reviewers routinely rubber-stamp agent outputs, their domain expertise and scrutiny deteriorate due to automation bias [Fügener et al., 2021, Goddard et al., 2012, Skitka et al., 1999, Parasuraman and Riley, 1997]. Consequently, the Verifiability Gap widens over time even if the technical system remains unchanged.

Corollary to P3 (Skipping oversight today makes governance harder tomorrow). A firm's human verification capacity will decline when it operates a high-autonomy system for longer without substantive human verification. This decline widens the Verifiability Gap even if the technical system remains unchanged.

6.2 Regulatory and Temporal Consequences

Historical reproducibility depends on preserved change, not time alone. Reproducing a past agentic decision requires more than merely resubmitting the original financial case. The verifier must reconstruct the exact historical execution environment: the specific model versions, system prompts, the settings that govern how the model chooses its words, retrieved data, tool responses, inter-agent messages, orchestration logic, and control rules. If a provider alters any of these components post-deployment and the original state was not explicitly pinned or escrowed, later re-execution tests a fundamentally different system.

Let j index system changes that occur before audit lag τ. Let μ_{dj} denote the materiality of change j to the complete decision chain for d, including propagated downstream effects. Let $p_{dj} = 1$ when the historical state affected by that change was preserved. Cumulative exposure to unpinned material change is:

$$M_d(\tau) = \sum_{j:t_j \leq \tau} \mu_{dj}(1 - p_{dj}), \qquad R_{dq} = r(M_d(\tau)), \qquad r(M + \Delta M) \leq r(M) \text{ for } \Delta M \geq 0. \tag{10}$$

Equation 10 assigns no direct effect to elapsed time. Audit lag (τ) degrades historical reproducibility only because it widens the window for unpinned material changes ($M_d(\tau)$) to accumulate. Because these changes are discrete such as a model release or API deprecation, reproducibility loss is typically abrupt rather than a smooth, continuous temporal decay [Pineau et al., 2021, Atıl et al., 2025].

P4 (System changes make old decisions harder to reproduce). Unpinned changes make past agentic financial decisions harder to reproduce. The likelihood of material reproduction will decrease with the cumulative number and materiality of those changes. Audit lag will reduce historical reproducibility only insofar as it increases exposure to such changes.

Serial agent governance requires chain-level preservation. In a serial architecture, the auditable object is not an isolated agent or the final output. It is the unbroken chain of causally material states. Seriality implies causal dependence: altering a transferred state alters a downstream risk assessment, control check, or final action. Consequently, every intermediate message, tool output, and approval signal becomes a critical component of the historical execution environment.

Preserving individual agents is useless if these intermediate handoffs are overwritten, compressed, or omitted. An unpinned upstream change propagates through every downstream agent, increasing μ_{dj} in Equation 10. Serial architectures thus multiply the number of states that must be perfectly preserved to retain R_{dq} and E_{dq}. By enabling broader delegated authority ($\rho_\sigma(A_d)$) while simultaneously multiplying the points of potential evidentiary failure, serial systems can severely widen the Verifiability Gap from both ends.

Corollary to P4 (Serial systems require chain-level preservation). Historical reproducibility will decrease with the number and materiality of causally dependent inter-agent

handoffs exposed to unpinned change. Preserving individual agents or final outputs will not be sufficient when intermediate states shape downstream financial actions.

Explainability requires faithful evidence of the agentic decision chain. In FinTech agentic systems, explainability requires more than a post hoc rationale for the final action. It depends on capturing the causally material states that actually drove the decision: model outputs, tool responses, inter-agent messages, and control decisions. In a serial architecture, a terminal agent might generate a fluent justification that completely masks a critical upstream error or altered state. Such a rationale fails to substantiate how delegated authority was exercised.

Let $C_{dq} \in [0,1]$ denote the coverage of causally material states in the retained record and $\phi_{dq} \in [0,1]$ its validated causal faithfulness. The actual explainability evidence retained is:

$$E_{dq} = C_{dq}\phi_{dq}, \qquad \frac{\partial E_{dq}}{\partial C_{dq}} = \phi_{dq}, \qquad \phi_{dq} = 0 \implies E_{dq} = 0. \tag{11}$$

Equation 11 demonstrates that expanding record coverage (C_{dq}) increases explainability only if those records are causally faithful (ϕ_{dq}). If faithfulness is zero, even exhaustive logging yields zero evidentiary value. Faithfulness is the vital link converting raw agentic states into substantive verification evidence [Gregor and Benbasat, 1999, Jacovi and Goldberg, 2020, Lanham et al., 2023, Turpin et al., 2023].

P5 (Agentic explanations count only when they track the decision chain). The retained explainability of a FinTech agentic decision depends on evidence across the entire decision chain. It increases with both the coverage and causal faithfulness of that evidence. It increases strictly in either dimension when the other is positive. A fluent terminal rationale adds no verification capacity for causally material upstream states that it omits or misrepresents.

6.3 Network-Level Consequences

End-to-end verification is bounded by the weakest handoff. Agentic financial decisions increasingly cross organizational lines. An upstream agent might supply a risk score or compliance check. A downstream institution then acts on it. This transfer is a "handoff." A handoff matters—it is causally material—if changing the transferred state, while holding other execution conditions fixed, changes the final decision.

Let h index these handoffs and s_h denote the transferred state. Let $Z_{dh}(s_h)$ denote the final downstream action under that state, holding all other execution conditions fixed. We define causal materiality as:

$$\chi_{dh} = \mathbf{1}\left\{\exists\, s_h \neq s_h' : Z_{dh}(s_h) \neq Z_{dh}(s_h')\right\}, \qquad \mathcal{H}_d^m = \{h : \chi_{dh} = 1\}. \tag{12}$$

Equation 12 separates trivial data transfers from causally material ones. A causally material handoff alters the final decision path. To verify the final action, every participant must save the exact data it received and the logic it applied to that data [Singh et al., 2019, South et al., 2025].

If the record of one causally material handoff is unavailable, the chain breaks. Downstream firms might still audit their own local steps. However, nobody can reconstruct the true basis of the final action. Serial chains therefore face a weakest-link constraint. The danger is not simply having more firms involved. The danger is that evidence must survive every causally material transfer.

Let $V_{dhq} \in [0,1]$ denote the verification capacity retained at handoff h. Let F_{dh} denote an evidentiary failure at that step. If B_{dn} is the end-to-end failure event across n causally material handoffs, then:

$$\begin{aligned} V_{dq}^{\mathrm{E2E}} &\le \min_{h \in \mathcal{H}_d^m} V_{dhq}, \qquad \min_{h \in \mathcal{H}_d^m} V_{dhq} = 0 \implies V_{dq}^{\mathrm{E2E}} = 0, \\ B_{dn} &= \bigcup_{h=1}^{n} F_{dh}, \qquad \Pr(B_{d,n+1}) - \Pr(B_{dn}) = \Pr(F_{d,n+1} \cap B_{dn}^{c}) \ge 0. \end{aligned} \tag{13}$$

Equation 13 establishes two rules. First, end-to-end verification cannot exceed the weakest causally material handoff. Second, adding more handoffs cannot reduce the chance of an audit failure. The increase is strict whenever the new handoff introduces a positive failure possibility not already contained in the existing chain.

> *P6 (One missing record breaks the whole chain).* In multi-firm agent networks, overall verification capacity is limited by the weakest causally material handoff. Additional handoffs create more potential points of evidentiary failure. They will therefore increase the risk of an end-to-end audit failure, even when individual firms audit their own local steps perfectly.

Shared providers create synchronized audit failures. Firms often appear to run independent AI systems. In reality, they frequently rely on the same shared model or API. When a provider updates or removes that shared tool, execution conditions change for everyone. This creates a shared risk. A single upstream update can prevent multiple firms from auditing their past decisions at the same time [Financial Stability Board, 2024, European Central Bank, 2024].

Relying on a shared provider is not inherently fatal. Firms can pin specific versions, escrow models, or build fallbacks. A shared audit vulnerability exists when firms cannot control historical replay.

Let $u_{ig} \in [0,1]$ denote institution i's reliance on shared component g. Let $\eta_{ig} \in [0,1]$ denote its control over historical replay. Its uncontrolled exposure is $e_{ig} = u_{ig}(1 - \eta_{ig})$. Let Z_g denote the random magnitude of a material upstream change and L_{ig} the resulting loss of reproducibility. If ω_i is the firm's network weight, then:

$$e_{ig} = u_{ig}(1-\eta_{ig}), \qquad L_{ig} = e_{ig} Z_g, \qquad L_g^{\mathrm{net}} = Z_g \sum_i \omega_i e_{ig}, \qquad \mathrm{Cov}(L_{ig}, L_{kg}) = e_{ig} e_{kg} \,\mathrm{Var}(Z_g) > 0, \tag{14}$$

for institutions i and k with positive exposure to the same component and $\mathrm{Var}(Z_g) > 0$. Equation 14 separates basic reliance from uncontrolled exposure. Reliance becomes an audit risk when institutions cannot replay past versions. As uncontrolled exposure grows, network-wide losses accumulate. Because one upstream change causes losses across every exposed firm, their audit failures become synchronized.

> *P7 (Shared models create shared audit failures).* Many financial institutions may rely on the same uncontrolled AI provider. In that setting, a material upstream update will create a larger and more synchronized loss of historical reproducibility across the network.

Table 4 summarizes the multilevel governance architecture developed across P1–P7.

Table 4: The multilevel governance architecture of the Verifiability Gap.

Prop.	Governance level and outcome	Mechanism	Observable implication
P1	Firm: delegated authority	Retained verification capacity limits defensible authority.	Granted authority plateaus below technical capability.
P2	Firm: fault escape	Unfaithful explanations weaken human review.	Action bounds outperform explanation-based review as faithfulness uncertainty rises.
P3	Firm: operating trade-off	Preserving evidence consumes speed, flexibility, and expertise.	Firms face a strict verifiability–efficiency frontier.
P4	Regulatory: historical fidelity	Material unpinned changes break continuity with the prior execution environment.	Historical-fidelity failures cluster around system updates, not elapsed time.
P5	Regulatory: evidentiary validity	Faithfulness turns compliance artifacts into substantive evidence.	Documentation can increase without improving actual auditability.
P6	Network: end-to-end verification	Every causally material handoff must preserve its own evidence.	One missing record prevents full end-to-end reconstruction.
P7	Network: common-mode exposure	Shared unpinned infrastructure changes many systems at once.	Historical-fidelity failures cluster across nominally independent firms.

7. Experiments on the Governance Mechanisms of the Verifiability Gap

The preceding theory defines the Verifiability Gap as a governance exposure. It arises when delegated authority exceeds the evidence retained to substantiate how that authority was exercised. We now examine controlled manifestations of the mechanisms that can produce this exposure. The experiments do not estimate the prevalence of the Verifiability Gap across financial institutions, nor do they directly test every firm-, regulatory-, and network-level consequence developed in Section 6. Instead, they isolate how provider change, execution-control loss, serial orchestration, trace divergence, output collapse, and model-version change can cause conventional audit indicators to overstate retained verification capacity.

The evidence comprises three linked studies. Studies 1 and 2 hold constructed financial cases fixed while varying provider release, execution controls, or orchestration. Study 3 holds two frozen versions of a transparent credit model against each other, testing whether current outcome reproducibility, historical outcome reproducibility, and faithful explainability can diverge within the same decision episode. Together, the studies ask not merely whether an answer repeats, but whether the institution retains evidence sufficient to substantiate the delegated authority that produced it.

7.1 Reproducibility Is a Governance Profile, Not a Scalar

Agentic governance must not collapse distinct phenomena into one misleading metric. We therefore decompose "reproducibility" into four explicit targets. They are *current outcome reproducibility, historical outcome reproducibility, material process reproducibility, and exact trace reproducibility.*

Let y_c^0 denote the historical financial action recorded for decision episode c at time t_0. Let $\mathbf{z}_c^0$ denote its recorded historical process trace. The trace includes materially relevant states, communications, tool invocations, and handoffs. Let $y_{c,k}(t)$ and $\mathbf{z}_{c,k}(t)$ denote the action and corresponding process trace obtained on repetition k under the environment available at time t. Throughout, $\mathcal{C}$ is the set of decision episodes under test, K the number of repetitions executed per episode, and $\mathcal{L}$

the set of actions the system is permitted to take.

Current outcome reproducibility. This metric operationalizes within-environment reproducibility: whether today's environment returns a stable action across repeated executions:

$$R_O(t) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{K} \max_{\ell \in \mathcal{L}} \sum_{k=1}^{K} \mathbf{1}[y_{c,k}(t) = \ell] \,. \tag{15}$$

High $R_O(t)$ establishes current stability across the observed repetitions. It does *not* prove that the system reproduces the historical decision.

Historical outcome reproducibility. This metric asks whether current re-execution recreates the exact historical action recorded at t_0:

$$R_H(t) = \frac{1}{|\mathcal{C}|K} \sum_{c \in \mathcal{C}} \sum_{k=1}^{K} \mathbf{1}\left[y_{c,k}(t) = y_c^0\right] . \tag{16}$$

In controlled studies, the baseline is an experimental condition rather than a single deployed decision. We therefore use the baseline modal verdict as a transparent proxy:

$$\widetilde{R}_H(t) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbf{1}[\widehat{y}_c(t) = \widehat{y}_c(t_0)] \,. \tag{17}$$

Here, $\widehat{y}_c(t)$ denotes the modal verdict across repetitions for case c at time t. We refer to $\widetilde{R}_H(t)$ as *baseline-modal historical fidelity.*

Material process reproducibility. This metric asks whether the materially relevant historical process can be reconstructed to satisfy the applicable evidentiary standard (σ):

$$R_P(t;\sigma) = \frac{1}{|\mathcal{C}|K} \sum_{c \in \mathcal{C}} \sum_{k=1}^{K} \mathbf{1}\left[\mathbf{z}_{c,k}(t) \equiv_\sigma \mathbf{z}_c^0\right] , \tag{18}$$

where $\equiv_\sigma$ denotes material equivalence. This standard is deliberately weaker than raw textual identity: a changed comma is immaterial, but a changed risk threshold is decisive.

Exact trace reproducibility. Modeling reproducibility specifically for multi-agent architectures requires capturing the complete, cascading chain of inter-agent communications, tool invocations, and state handoffs. We formalize this diagnostic as *exact trace reproducibility* (R_T). It asks whether the entire multi-agent execution sequence reproduces perfectly across independent runs:

$$R_T(t) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbf{1}[\mathbf{z}_{c,1}(t) = \cdots = \mathbf{z}_{c,K}(t)] \,. \tag{19}$$

If $R_T(t) < 1$, the multi-agent decision process is not exactly reproducible across all observed runs. Benign textual variation, such as a rephrased internal summary, does not automatically constitute a material governance failure under $\equiv_\sigma$. We therefore treat exact trace reproducibility as a strict diagnostic of within-environment process stability. It is not a direct substitute for R_P.

Verdict differentiation as a degeneracy diagnostic. High $R_O(t)$ is highly misleading if an agent simply defaults to the same blanket action for every case. We therefore track $D_V(t)$, the normalized within-family verdict spread. This measure diagnoses output collapse. When $R_O(t)$ rises while $D_V(t)$ falls for a case family that requires differentiated verdicts, the system has lost its ability to distinguish cases. The resulting stability is degenerate.

The true empirical object is therefore a profile:

$$\mathcal{R}(t;\sigma) = \big(R_O(t), R_H(t), R_P(t;\sigma), R_T(t); D_V(t)\big). \tag{20}$$

The semicolon separates the four reproducibility targets from $D_V(t)$, which serves as a degeneracy diagnostic rather than an additional reproducibility target.

Three operational patterns are especially consequential for governance:

- $R_O = 1$ *but* $R_H < 1$: The current system is perfectly stable, but historical decision recovery is incomplete.
- R_O *is high but* $R_T = 0$: Stable final decisions coexist with trace-level non-reproducibility across every case, even though the material governance significance of that variation still depends on R_P.
- R_O *rises while* D_V *falls*: Apparent reproducibility is degenerate when the increase in stability is achieved by collapsing cases requiring differentiated assessment toward the same blanket action.

The theoretical implication is fundamental: *reproducibility is a composite governance profile, not a scalar model property.*

The profile is not merely prudent. Two results establish that no single within-case statistic can substitute for it; proofs are in Appendix S2.

Theorem 5 (Range mismatch)**.** *Let $L = |\mathcal{L}|$ be the number of actions the system is allowed to take, m the number of model calls in one decision, and $\bar{p} < 1$ an upper bound on the probability that any single call repeats its own output. Then $R_O \geq 1/L$ always, whereas $R_T \leq \bar{p}^m \to 0$ and, by Theorem 1, $I(C;Y) \to 0$ geometrically in depth. As depth grows, the attainable region of $(R_T, I(C;Y), R_O)$ therefore approaches $\{0\} \times \{0\} \times [1/L, 1]$: the process record and the case information are driven to zero while the outcome metric remains bounded away from zero.*

In plain terms: the process record can fall to nothing while the outcome metric stays high. This is stronger than saying outcome agreement can mislead: the metric is bounded below by $1/L$ on a scale whose failure mode lies at zero, so its range does not contain the failure it is being used to exclude.

Corollary 4 (No threshold on R_O is a sufficient audit criterion)**.** *For any rule of the form "authorize while $R_O \geq \tau$" there exist configurations that satisfy it with $R_T = 0$ and $I(C;Y) = 0$. The constant policy $Y \equiv \ell_0$ attains $R_O = 1$ with $I(C;Y) = 0$.*

Theorem 6 (Degeneracy is not identifiable within cases)**.** *Let $\mathcal{A}$ return the case-appropriate action $\ell^\star(c)$ with probability one, and let $\mathcal{B}$ return a fixed ℓ_0 with probability one. Both induce a point-mass verdict distribution for every case. Every statistic that is a function of the within-case verdict distributions alone therefore takes the same value under $\mathcal{A}$ and $\mathcal{B}$, for every number of repetitions K.*

In plain terms: a system that always answers correctly and a system that always answers the same thing look identical if you only ever re-run one case.

Corollary 5. *Separating $\mathcal{A}$ from $\mathcal{B}$ requires at least two cases whose correct actions differ, and a statistic comparing modal verdicts* across *those cases. Decision families with differentiated correct actions, together with D_V, constitute the minimal design that identifies degeneracy.*

Remark 2. R_O is a functional of the within-case distributions; D_V and $I(C;Y)$ are functionals of the cross-case distribution of modal verdicts. R_O is not a function of D_V, and no claim here asserts otherwise. Theorem 6 states why the second family is not redundant: the first is blind to degeneracy by construction, at any sample size.

7.2 Experimental Design and Identification Strategy

Studies 1 and 2 share a common test bed of 32 constructed financial cases (eight each in credit, trading, AML, and portfolio rebalancing). Within each case, the financial objective, facts, instructions, and allowable actions are fixed. The design isolates specific governance mechanisms that can erode reproducibility; it does not estimate average prevalence across the financial sector.

Release-timeline arm: changing the object of audit. This arm tests whether a current provider release reproduces actions generated by an earlier release. We execute the 32 cases across four dated releases of Anthropic's Claude Opus line over 185 days (three repetitions per cell). The releases are `claude-opus-4-5` (24 November 2025), `claude-opus-4-6` (4 February 2026), `claude-opus-4-7` (14 April 2026), and `claude-opus-4-8` (28 May 2026). Comparing later modal verdicts to the earliest-release baseline yields baseline-modal historical fidelity, $\widetilde{R}_H(t)$. This design captures the effect that an institution encounters when a provider updates the system. Such an update may bundle changes to model behavior, capabilities, and infrastructure. We do not attribute an observed verdict shift to any particular undocumented internal change.

Decoding-control arm: changing the means of replay. This arm isolates the execution controls required for stable replay. Two provider-exposed settings govern that capability. *Temperature* scales how much randomness the system is permitted when it selects each successive word. At zero it always takes its single most likely continuation. The same input then tends to return the same text. As temperature rises, the same input can return different text on each execution. A *random seed* fixes the draw used when that randomness is exercised, so that a repeated run samples identically. Neither setting is the institution's to keep. Both are exposed, or withdrawn, by the provider. We use a fixed, locally served `llama3.2:3b` model. We examine four conditions: (1) temperature zero, fixed seed; (2) temperature zero, varied seed; (3) no temperature sent, varied seed; and (4) neither sent. The first condition is an implementation check; the main-text comparison reports the three conditions that do not fix the seed. Comparing Conditions 2 and 3 isolates the loss of temperature control while seed availability is retained; comparing Conditions 3 and 4 identifies the additional loss associated with seed unavailability. Because the executable is fixed, this arm measures current outcome reproducibility, $R_O(t)$.

Execution-environment arm: changing where the same call runs. This arm holds the cases, the prompts, the model version, and the decoding settings fixed, and varies only whether execution occurs on a locally served model or on Anthropic's hosted Claude endpoint. Ten repetitions per case are executed in a single session, so no release change can intervene. The arm measures R_O under the most favorable replay conditions an institution can arrange.

Orchestration arm: changing the governance architecture. Study 2 isolates the decision topology by sweeping configurations from one to fifty agents. The multi-agent setup strictly separates information access, analysis, and authority (Figure 5). Upstream specialists pass intermediate reports along a fixed graph, and only a terminal Decider issues the final verdict. Every intermediate record is retained. This allows us to locate the first observed divergence and trace how it propagates. The design yields current outcome reproducibility $R_O(t)$, exact trace reproducibility $R_T(t)$, and verdict differentiation $D_V(t)$.

Adding agents widens the graph more than it lengthens it, and the distinction matters for the theory. Every layer holds three specialists, each reading two reports from the layer above, so

a configuration of fifty specialists is 52 nodes arranged in 19 sequential stages (Figure 6). The exponent in the contraction bound of Theorem 1 is the number of sequential stages. For the largest configuration that exponent is 19, not 50. We report agent counts on the horizontal axis of every sweep figure because that is the quantity a firm chooses; the corresponding stage counts are 3, 4, 6, 9, 12, 16, and 19.

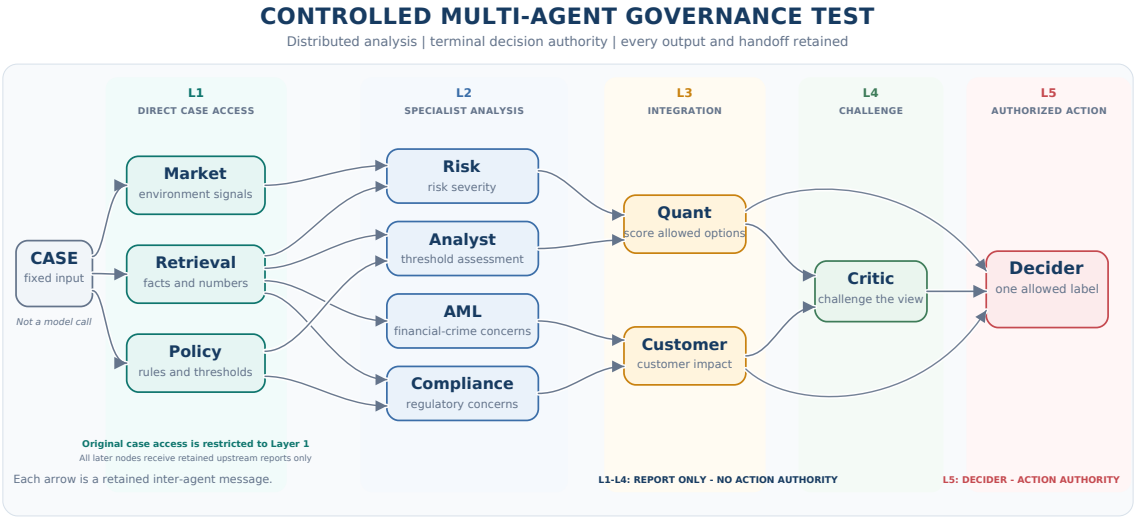


Figure 5: Controlled multi-agent governance test. Only Retrieval, Market, and Policy observe the original financial case; downstream roles observe only the records transmitted along designated edges. Specialist roles are report-only, and only the terminal Decider can issue the final financial verdict. Retaining every role output and handoff makes it possible to locate the first observed divergence and trace how it propagates through the governance chain.

Identification boundary and triangulation. To cleanly isolate orchestration, the communication graph is fixed ex ante, tool payloads are frozen, and the graph is traversed once without iterative self-correction.

These four arms are not pooled because they target different evidentiary failures. The release arm changes the executable ($\widetilde{R}_H$); the decoding arm narrows the replay controls (R_O); the execution-environment arm moves the same call between local and hosted infrastructure (R_O); and the orchestration arm alters the decision topology (R_O, R_T, D_V). Their convergence constitutes mechanism triangulation, not a single pooled treatment effect. Study 3 is analytically separate and detailed in Section 7.5.

7.3 Study 1: Provider Dominance Over the Governance Environment

Our first study asks whether an institution can preserve historical decision behavior after a provider changes the system. We examine changes to both the model version and the execution controls available for replay. The release-timeline test holds 32 financial cases fixed across four dated Claude Opus releases. A separate decoding-control grid holds the cases and model fixed while varying temperature and seed availability. A third, paired test holds everything fixed and moves the same call between local and hosted execution. These tests are not pooled into a single causal estimate. Instead, they identify three distinct channels through which retained reproducibility depends on a provider-controlled audit surface.

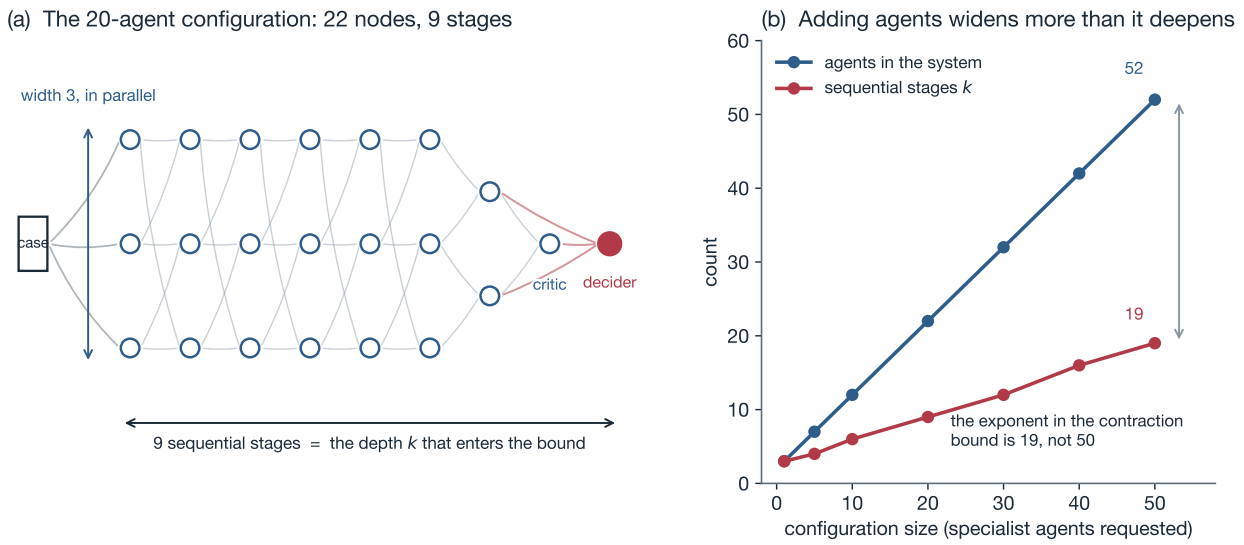


Figure 6: Agent count and chain depth are different quantities. **(a)** The twenty-agent configuration. Three first-layer specialists read the case; every later specialist reads two reports from the layer above and never sees the case itself. A critic challenges the emerging view, and only the terminal Decider issues an action. The graph is three agents wide at every layer, so twenty-two nodes span nine sequential stages. **(b)** Across the sweep, node count grows linearly with the configuration size while the number of sequential stages grows far more slowly. The largest configuration has 52 nodes but 19 stages, and it is the stage count that enters the geometric contraction bound.

7.3.1 Silent Risk Shifts and the Erosion of the Audit Surface

We examine four dated releases of Claude Opus over 185 days. Baseline-modal historical fidelity ($\widetilde{R}_H$) remained perfect (1.000) through the second release. It then dropped to 0.906 in the third and fourth releases (Figure 7a). In both of these later releases, 29 of the 32 modal decisions matched the baseline. Across the releases as a whole, however, five distinct cases changed at least once.

Every deviation moved toward a more conservative action. Four trading cases shifted from *buy* to *hold*, and one rebalancing case shifted from *rebalance* to *hold*. (Credit and AML cases remained stable.) This is not merely generic "model drift." It is a silent shift in the operative risk posture. Absent release-specific revalidation, an institution can easily continue operating a system whose behavioral policy has shifted without internal authorization.

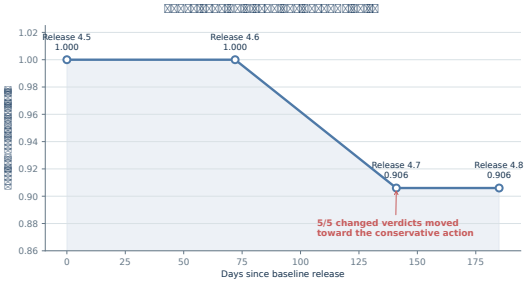
Stable aggregate metrics also mask underlying case churn. The fidelity score remained flat at 0.906 across both later releases, yet the specifically affected cases differed. Of the three cases that diverged from baseline in the third release, one remained changed in the fourth, and two reverted to baseline. Two entirely different cases then changed for the first time.

A flat aggregate fidelity score therefore fails to identify a stable remediation population. For governance, this means an institution cannot determine from portfolio-level stability alone which specific historical decisions require review.

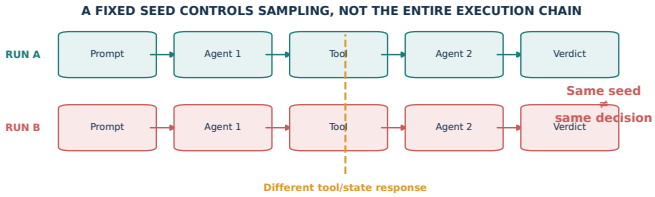
The more consequential finding concerns the audit surface itself. In later releases, Anthropic withdrew institutional control over `temperature`—an execution parameter critical for stable replay. The withdrawal is total rather than partial. A request to `claude-opus-5` carrying `temperature` returns `HTTP 400`, "`'temperature' is deprecated for this model`," and the same applies to `top_p` and `top_k`. This holds for `temperature` = 0, the setting an auditor would use. The line has never exposed a random seed at all. The provider thus altered not only the system being

audited, but also the very controls required to conduct the audit. We term this *control-surface dependence*: an institution's retained reproducibility relies on execution controls it does not own and cannot unilaterally preserve.

A fixed random seed cannot compensate for this missing environment. A seed only constrains sampling conditional on the exact same model, instructions, tool responses, and upstream state. If any upstream component changes, the seed applies to a fundamentally different conditioning context (Figure 7b).



(a) Release-timeline drift.



(b) Limits of seed-based replay.

Figure 7: Provider dominance over the audit surface. **(a)** Baseline-modal historical outcome drift across four dated Claude Opus releases. The experimental proxy $\widetilde{R}_H$ remains 1.000 through the second release, falling to 0.906 in the third and fourth. Every changed verdict moves toward a more conservative action. Because successive releases also alter capabilities, this captures release-timeline drift rather than silent same-version drift. **(b)** A fixed seed controls sampling, not the complete execution chain. It cannot recreate a historical decision if the model state, tool responses, or intermediate context drift.

A separate controlled experiment isolates these execution controls directly. With temperature pinned to zero, both one- and ten-agent configurations achieved perfect current outcome reproducibility (R_O) across varied seeds. Removing the temperature control reduced reproducibility; removing the seed reduced it further (Table 5). In this setting, temperature was the primary control for verdict stability, while seed availability provided an additional stabilizing effect once temperature control was absent. While this experiment and the provider observation use different model settings, they demonstrate the same theoretical dependency: reproducibility is strictly bounded by the control surface exposed to the institution.

Table 5: Execution controls and current outcome reproducibility. Same model, same 32 cases, and five repetitions per case; entries are verdict-level estimates of R_O.

Temperature	**Seed**	**1 agent**	**10 agents**
Pinned to 0	Varied	1.000	1.000
Not sent	Varied	0.912	0.794
Not sent	Not sent	0.831	0.744

7.3.2 Pinned Controls Restore Reproducibility Locally but Not on a Hosted Endpoint

Table 5 invites an obvious remedy: pin temperature to zero, freeze the model version, and the problem appears to dissolve. Three paired arms show that this remedy works on a locally served model, fails on a hosted one, and is not even available on a frontier one. We first executed the

same 32 cases ten times each at temperature zero against two execution environments. The first is the locally served `llama3.2:3b` used elsewhere in Study 1. The second is Anthropic's hosted endpoint running `claude-haiku-4-5-20251001`, pinned to that dated release. Stimuli, prompts, and decoding settings were identical across arms. Only the execution environment differed.

The local arm reproduced itself exactly, returning the modal verdict in all 320 executions. The hosted arm returned it in 319 (Table 6). The single deviation fell on a borderline credit case with a subprime credit score, a debt-to-income ratio of 47 percent, and one prior delinquency. Ten identical executions returned *deny* nine times and *refer* once.

A third arm removes the remaining controls entirely. On `claude-opus-5` the institution cannot pin temperature, cannot pin a dated release, and has never had a seed, so the tightest available configuration is simply the provider default. We executed the same 32 cases 30 times each in a single session, 960 executions. The modal verdict returned in 959. The deviation is instructive, because it did not fall on a marginal application. The case carries a credit score of 754, a debt-to-income ratio of 31 percent, and an income roughly five times the amount requested, against one delinquency in 24 months. Twenty-nine executions approved it. One referred it to a human underwriter, reasoning that "the recent 24-month delinquency warrants human review of its severity and cause," where the other twenty-nine had reasoned that ample income "outweighs the single isolated delinquency." The irreproducibility is therefore not noise that a filter could remove. It is a defensible alternative judgment on identical facts, surfacing at random.

Table 6: Within-session reproducibility under the tightest replay controls each endpoint exposes. Same 32 cases and identical prompts throughout; the arms differ in the controls available, not in the stimulus. The commercial models offer an optional internal deliberation step before answering; it is switched off here, so all three arms answer directly, as the local model does.

Environment	**Model**	**Replay controls exposed**	**Modal verdict**	**Cases differing**
Local, single stream	`llama3.2:3b`	temperature 0; version pinned	320 / 320	0 / 32
Hosted, shared	`claude-haiku-4-5`	temperature 0; dated release	319 / 320	1 / 32
Hosted, shared	`claude-opus-5`	none: provider default, undated alias	959 / 960	1 / 32

We report these as counterexamples and not as estimates. One deviating execution does not support a rate, and a difference of one case in 32 is not distinguishable from chance on this sample. The three arms also differ in repetition count and in the controls available, so their execution-level figures are not three measurements of one quantity. What the deviations establish is narrower and sufficient: reproducibility is not guaranteed on a hosted endpoint, either when every exposed control is pinned or when no control is exposed at all. The paired local control shows that the failure is not a general property of language models, since the identical protocol reproduced perfectly on local execution. Both hosted deviations landed on credit cases, and both moved toward *refer*, the action that transfers the decision to a human. Control-surface dependence therefore does not end when the institution pins every setting the provider exposes. Some of the variation that governs a financial action lives in execution infrastructure that exposes no control at all.

7.4 Study 2: Orchestration, Serial Dependence, and Output Collapse

The second study holds the locally served model, the thirty-two financial cases, frozen tool payloads, and observed execution settings fixed. It changes only the agentic architecture. The one-versus-ten-

agent comparison isolates the initial move from a single call to a serial pipeline. The one-to-fifty-agent sweep then tests three possibilities: monotonic decay, trace instability, or apparent recovery through output collapse. We measure current outcome reproducibility, exact trace identity, and verdict differentiation separately.

7.4.1 Architecture Is Policy: Orchestration Changes Decisions Even When Each Architecture Is Reproducible

We changed only the orchestration, moving from one call to the multi-agent pipeline. Mean current outcome reproducibility fell from 0.903 to 0.747 (Figure 8a). The family-level pattern is more revealing than the mean. Credit, trading, and rebalancing became less reproducible, whereas AML rose from 0.850 to 0.988 because the multi-agent system nearly always escalated the case. Constant action increased the metric.

Orchestration also changed the decisions themselves. In the temperature-zero condition, both architectures were perfectly reproducible within their own execution environments, yet their modal verdicts matched on only 16 of 32 cases. Thus, perfect within-system reproducibility does not imply policy invariance across architectures. Agent roles, topology, aggregation, and the information supplied to the terminal Decider are not neutral implementation details. They form a latent policy layer that can alter the institution's operative risk posture without an explicit policy decision. The designed roster partly conditions the direction of this shift. The evidence therefore supports a structural effect of architecture, not a universal claim that multi-agent systems always become more conservative.

7.4.2 Trace–Outcome Decoupling: The Verdict Can Reproduce While the Recorded Path Changes

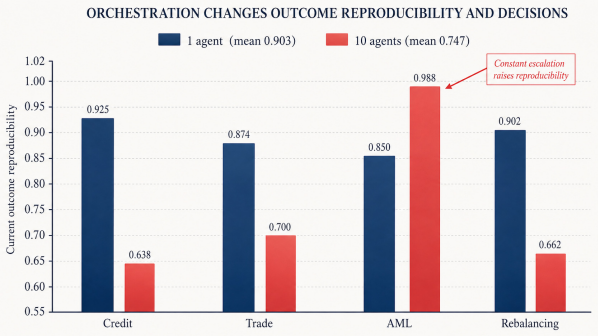
Outcome agreement also overstates the stability of the supporting record. In the multi-agent condition, the recorded output at every stage differed across repeated executions in all 32 cases, beginning with the first-layer retrieval record. The final verdict differed in only 25 of 32 cases (Figure 8b). Downstream agents therefore absorbed or compressed some upstream variation before it reached the final action. The organization can observe a stable verdict even though the recorded path to that verdict is nonidentical.

This pattern also rejects a naive independent-compounding interpretation. Variation enters early, is transformed through causally dependent handoffs, and is often compressed by the terminal categorical decision. The experiment directly establishes exact trace nonidentity, not that every linguistic difference is materially non-equivalent under every evidentiary standard. Its governance implication is nevertheless decisive: terminal verdict agreement cannot be used as proof that the material process has been reproduced. That inference requires a separate $R_P(t;\sigma)$ test.

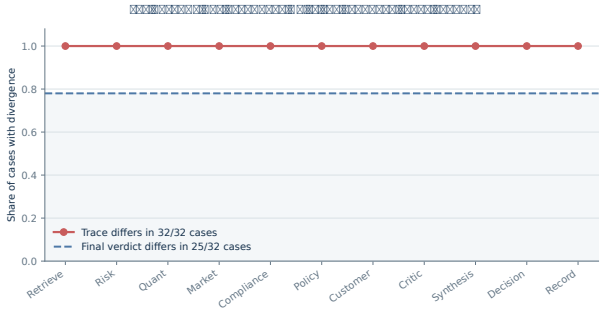
7.4.3 Degenerate Outcome Reproducibility: A System Can Look More Stable Because It Has Stopped Distinguishing Cases

Scaling the system from one up to fifty agents reveals a stark, non-monotonic pattern (Figure 9). Current outcome reproducibility (R_O) drops sharply from 0.912 at one agent to 0.750 at five agents, before fluctuating between 0.750 and 0.875.

This apparent recovery at larger scales is an illusion. It is not accompanied by a stable process record. Across all seven architectural configurations and 1,120 total executions, not a single case produced an exactly identical full set of agent outputs (Table 7).



(a) Architecture-level reproducibility.



(b) Trace–outcome decoupling.

Figure 8: Orchestration changes both the decisions being reproduced and the records supporting them. **(a)** Mean R_O falls from 0.903 in the one-agent configuration to 0.747 in the ten-agent configuration. AML moves in the opposite direction because near-constant escalation makes the final verdict artificially easy to reproduce. **(b)** Recorded traces differ in all 32 cases and at every stage, while the final verdict differs in only 25 of 32 cases. The comparison establishes exact trace nonidentity; establishing material process equivalence requires a separate evidentiary test under $\equiv_\sigma$.

Instead, the recovery stems from a severe collapse in verdict differentiation. At one agent, the eight AML cases cleanly split into three *clear* and five *escalate* decisions. By forty agents, the system blanket-escalated all eight. The credit family degraded similarly, shifting from a differentiated split (seven *approve*, one *deny*) to near-uniformity (seven *deny*, one *refer*). At forty agents, three of the four case families applied the exact same blanket verdict to all eight of their underlying cases.

The normalized within-family verdict spread (D_V) plummeted to 0.086 at forty agents and flatlined there. The share of decisions that both successfully reproduced *and* meaningfully distinguished between cases collapsed from 30% at one agent to just 7% at fifty agents.

We term this *degenerate outcome reproducibility*. The metric R_O superficially improves only because the system's action distribution collapses into blanket defaults. If a firm monitors outcome reproducibility in isolation, it might perversely rank the fifty-agent pipeline above the ten-agent one. It would systematically reward a credit model for blanket denials or an AML system for universal escalation. The metric falsely signals robust auditability at the exact moment the system loses the core discriminatory capability it was authorized to exercise.

Why the apparent recovery occurs. The mechanism starts with the base model, not orchestration. The sweep used an environment in which temperature was not supplied and seeds varied. Even the single-agent `llama3.2:3b` benchmark failed exact trace replay (0/32 matches) and yielded an R_O of 0.912. The base LLM injects textual variation before any inter-agent handoff occurs. Pinning temperature to zero restored perfect verdict reproduction at this model scale. That restoration is scale-dependent, and it should not be read as a general property of the control. Replicating the same arm on `llama3.1:8b` leaves current outcome reproducibility short of one at ten agents. Pairwise trace identity in that configuration is .45 on `llama3.1:8b`, against .94 on `llama3.2:3b` under the identical condition (Table S3). The control that looked sufficient at 3B is only partially effective at 8B.

The non-monotonic curve is therefore not a contest between two forces. One of the two is already saturated before the sweep begins. Exact trace reproducibility is zero in *every* configuration and at both model scales, the single-call configuration included. No case repeated its full record even once, in 1,120 executions per model. *Stochastic expansion* — the accumulation of textual variation

Table 7: Agent count, current outcome reproducibility, exact record identity, and verdict differentiation. Every configuration uses the same locally served `llama3.2:3b` model on the same machine, the same 32 cases, and five repetitions per case; temperature is not supplied and seeds vary across repetitions. The one-agent row is therefore the reproducibility benchmark of the underlying LLM under the same execution condition, rather than a different control arm. Exact graph structure and call accounting for each scale are generated from the replication configuration files.

Agents	R_O	**Exact record identity**	**Cases with a verdict split**	D_V
1	0.912	0/32	13/32	0.324
5	0.750	0/32	25/32	0.336
10	0.794	0/32	21/32	0.367
20	0.869	0/32	13/32	0.264
30	0.800	0/32	21/32	0.286
40	0.875	0/32	14/32	0.086
50	0.850	0/32	18/32	0.086

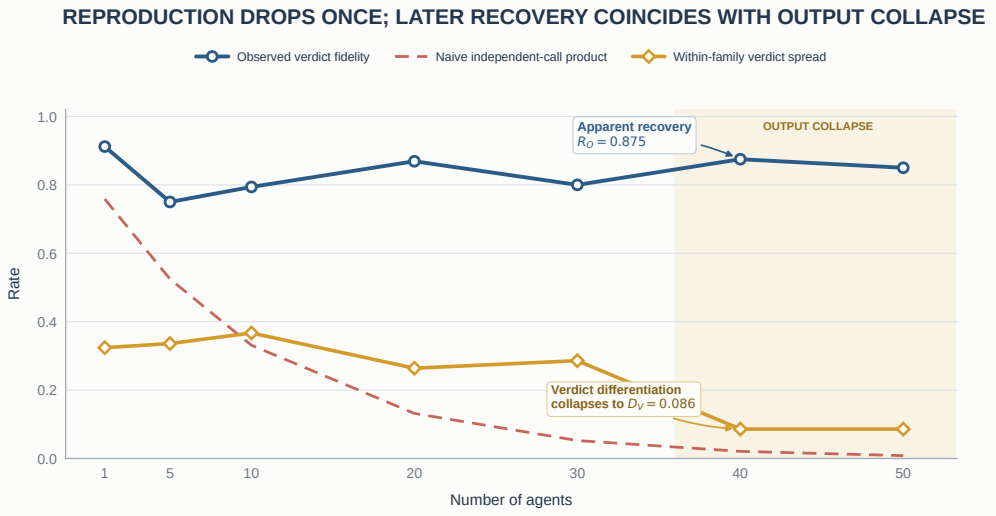


Figure 9: Why outcome reproducibility can recover while the underlying process remains nonreproducible. The same locally served `llama3.2:3b` model is used throughout. Moving from one to multiple agents initially adds opportunities for variable LLM outputs to enter the execution trace, reducing R_O. At larger scales, repeated summarization and the final categorical decision bottleneck map increasingly nonidentical traces onto a narrow set of default actions; R_O therefore partially recovers as D_V collapses. The dashed benchmark assumes that call-level differences propagate independently to the verdict and thus misses this many-to-one compression mechanism.

across added calls — therefore has no room left to operate on this metric. It reached its floor at one agent.

What remains is *categorical compression*, and it accounts for the entire curve. With $R_T = 0$, Lemma 1(ii) makes the outcome metric equal the compression probability exactly. Every movement in R_O across the sweep is therefore movement in κ. A serial pipeline is not simply an independent product of call-level error rates. Downstream agents repeatedly summarize nonidentical upstream text, and the final Decider maps a high-dimensional execution trace into a low-dimensional action

space. This process absorbs immaterial wording differences, preventing multiplicative decay. However, at extreme scales, absorption becomes overcompression. The system erases vital case facts needed to distinguish materially different cases. Our roster explicitly surfaces risk and compliance concerns. Consequently, conservative defaults such as *deny* or *escalate* become the dominant absorption states.

The mechanism replicates at a larger scale, attenuated. The mechanism holds at a larger scale. We repeated the full sweep on `llama3.1:8b`: same code, same thirty-two frozen cases, same seeds, same seven configurations, 1,120 executions. Figure 10 places the two scales side by side. Three things recur exactly. Exact trace identity is zero in every configuration. The local maximum of R_O falls at forty agents in both. The collapse in verdict differentiation falls at forty agents in both.

The magnitudes differ, and in the direction that matters for governance. At forty agents the eight-billion-parameter model retains 2.5 times the differentiation of the three-billion-parameter one ($D_V = 0.214$ versus 0.086), so it does not enter the range in which a decision family has stopped distinguishing its cases. It nevertheless loses 59% of its own differentiation there, and its outcome reproducibility is lower than the smaller model's at every configuration beyond ten agents. Between these two local models, scale moderates degeneracy without removing it.

A frontier model reproduces its actions better and its record no better. We extended the comparison to a commercial frontier model. Three configurations—one, ten, and forty agents—were executed on `claude-sonnet-5` with three repetitions per case, 5,310 calls in total. The provider does not disclose a parameter count. The system is a commercial model at the top of its provider's range, and we take it to exceed 10^{11} parameters, at least an order of magnitude beyond either local model. Nothing below depends on the exact figure. Because this arm uses three repetitions, we recompute the local results on their first three repetitions, so every comparison is at matched k.

Three results follow, and they do not point the same way. First, exact trace reproducibility is zero in every configuration for this model as well. No execution record repeated itself, at any scale tested. Second, outcome reproducibility is *higher* than the eight-billion-parameter model's, not lower: 0.958, 0.958 and 0.938 against 0.844, 0.802 and 0.802. The advantage holds case by case, not merely on average (sign tests over the 32 matched cases: $p = .007$ at one agent, $p < .001$ at ten, $p = .004$ at forty). Third, verdict differentiation starts far higher and falls further. It begins at $D_V = 0.749$, more than double either local model, and reaches 0.239 at forty agents, a loss of 68%. Because only three configurations were executed for this model, that loss is measured from the largest value observed among them and is therefore a lower bound.

These results revise an earlier generalization rather than the theory. The contraction bound of Theorem 1 is stated in sequential stages, not in model size, and it never predicted that outcome reproducibility would fall with capability. What the frontier model demonstrates is a separation. It reproduces its own actions more often than a small local model. It does not reproduce its own record any more often, since that quantity is zero at every scale. It does not retain more of its differentiation under orchestration, since it loses a comparable share. Capability purchases a higher starting point. It does not purchase auditability.

Two cautions bound these comparisons. The three endpoints do not share a decoding condition, because temperature is settable on the local models and rejected by the commercial endpoint, so part of the outcome-reproducibility difference may reflect default sampling rather than capability. And D_V is unstable at these sample sizes: recomputing the eight-billion-parameter model on three repetitions moves its forty-agent value from 0.214 to 0.414. We therefore report change within a model, which is stable, and avoid ranking models against one another at a single configuration.

The serial system appears to outperform the naive independent-call product for a perverse reason. It maps wildly divergent decision paths to the exact same terminal action. Let $T_{n,r}$ denote

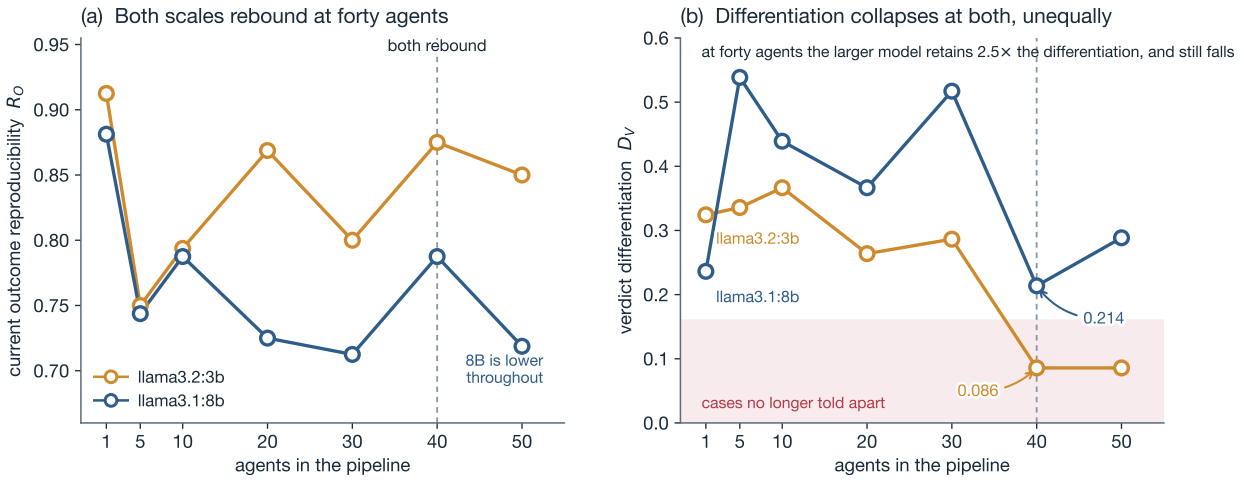


Figure 10: The same sweep at two model scales. **(a)** Current outcome reproducibility. Both scales fall, then recover at forty agents; the larger model is lower throughout beyond ten agents. **(b)** Verdict differentiation. Both collapse at forty agents. The smaller model enters the range in which cases within a decision family are no longer told apart; the larger model retains 2.5 times the differentiation and stays above it, while still losing most of its own. Exact trace reproducibility is zero at every point on both curves, so by Lemma 1(*ii*) panel (a) is the compression probability κ throughout.

the complete trace produced in repetition r of an n-agent configuration, and let $Y_{n,r} = g_n(T_{n,r})$ denote the final verdict. The system falls into degenerate reproducibility when:

$$T_{n,r} \neq T_{n,r'} \qquad \text{but} \qquad g_n(T_{n,r}) = g_n(T_{n,r'}) = y^\dagger, \tag{21}$$

for a growing share of cases, where $y^\dagger$ is a blanket default action. The exact same information loss that prevents multiplicative decay drives D_V toward zero, creating the mathematical illusion of a highly reproducible system.

7.5 Study 3: A Reproducible Credit Model Can Still Fail to Reproduce History

The earlier release study produced no verdict changes among its constructed credit cases. That null result is specific to those cases and releases; it should not be mistaken for immunity of credit decisions to historical- verification failure. To demonstrate the mechanism directly in a stable credit setting, we conduct a preserved-version experiment using public data and fixed decision thresholds. By separating this numerical test from the multi-agent pipelines, we remove generative stochasticity and inter-agent orchestration as necessary explanations for the result.

Because both model versions are preserved in the experiment, the design also distinguishes the failure condition from its governance remedy. Replaying the current version against a decision produced by the historical version tests historical fidelity. Replaying the preserved historical version tests whether retaining the original executable restores the capacity to reproduce and substantiate the decision.

Design, data, and decision rule. We use the public FICO Home Equity Line of Credit (HELOC) challenge dataset [FICO, 2018]. To keep explainability directly testable, we fit standardized logistic models using five core credit-report variables, yielding 9,682 complete records. We fit version 1 (v_1)

on 5,808 records. We then fit version 2 (v_2) on an expanded estimation set. That set includes the original 5,808 records and 1,937 additional records. Both versions use the same variables, model architecture, estimation code, and decision rule. We evaluate both versions on the same strict 1,937-record holdout.

The models estimate the probability of adverse credit performance, denoted p. The illustrative policy approves cases when $p < .45$, refers cases for substantive human review when $.45 \leq p < .55$, and denies cases when $p \geq .55$. These thresholds are chosen to expose the verification mechanism, not to prescribe an underwriting policy.

Aggregate validation masks changed financial actions. At the portfolio level, the two versions appear nearly identical: $\text{AUC}_{v_1} = .7901$ and $\text{AUC}_{v_2} = .7899$. Yet AUC and historical action fidelity evaluate different objects. AUC is a threshold-independent measure of how well a model ranks cases. A financial action is a threshold-dependent, case-specific consequence of that score. A model refresh can therefore preserve nearly the same overall ranking while moving individual cases near a policy boundary from one action category to another.

That is precisely what occurs. Although aggregate discrimination is virtually unchanged, 23 of the 1,937 held-out policy actions change after the refresh, yielding held-out historical fidelity of

$$R_H(v_2 \leftarrow v_1) = \frac{1{,}914}{1{,}937} = .9881. \tag{22}$$

The result is not a contradiction. The refresh changes little about aggregate ranking performance, but small score movements become financially consequential when they cross a discrete policy threshold. Similar aggregate performance therefore does not guarantee historical action fidelity.

The decision-level verification gap. Consider source-row index 2270, one of the 23 discordant cases, selected to make the threshold-crossing mechanism directly observable. Version v_1 estimates $p = .5580$ and denies the application. Version v_2 estimates $p = .5488$ and refers the same application. The score changes by only .0092, but that movement crosses the fixed denial threshold and changes the financial action.

Across ten repeated executions, each frozen version perfectly returns its own version-specific score and action. Thus:

$$R_O(v_1) = R_O(v_2) = 1, \qquad R_H(\text{case } 2270) = 0. \tag{23}$$

Equation 23 captures the central separation. Current outcome reproducibility is a within-version property: it asks whether the same frozen executable returns the same action again. Historical fidelity is a cross-version property: it asks whether the current executable reproduces the action taken by the historical executable. Repeated execution of v_2 perfectly reproduces the current referral, but it never recovers the historical v_1 denial. Perfect current reproducibility can therefore coexist with complete historical-fidelity failure at the decision level.

Position, not magnitude, decides which actions change. Case 2270 is representative rather than exceptional. An action can change only when the score movement $|\Delta p|$ exceeds the distance from the historical score to the nearest policy cutoff. All 23 discordant cases lie above that boundary and none falls below it, as panel (a) of Figure 11 shows. The condition is necessary but not sufficient, because a score can also move away from a cutoff rather than across it. The score movements themselves are unremarkable. Their median is .0088, only 2.6 times the .0034 median across all held-out cases. What sets the 23 apart is where they started. Every one begins within .0161 of a cutoff, the 6.9th percentile of the held-out distance distribution. The finding is not an artifact of

the .45 and .55 cutoffs. Across the 39 admissible cutoff pairs on a .05 grid, every pair changes at least 14 held-out actions, with a median of 21, as panel (b) reports.

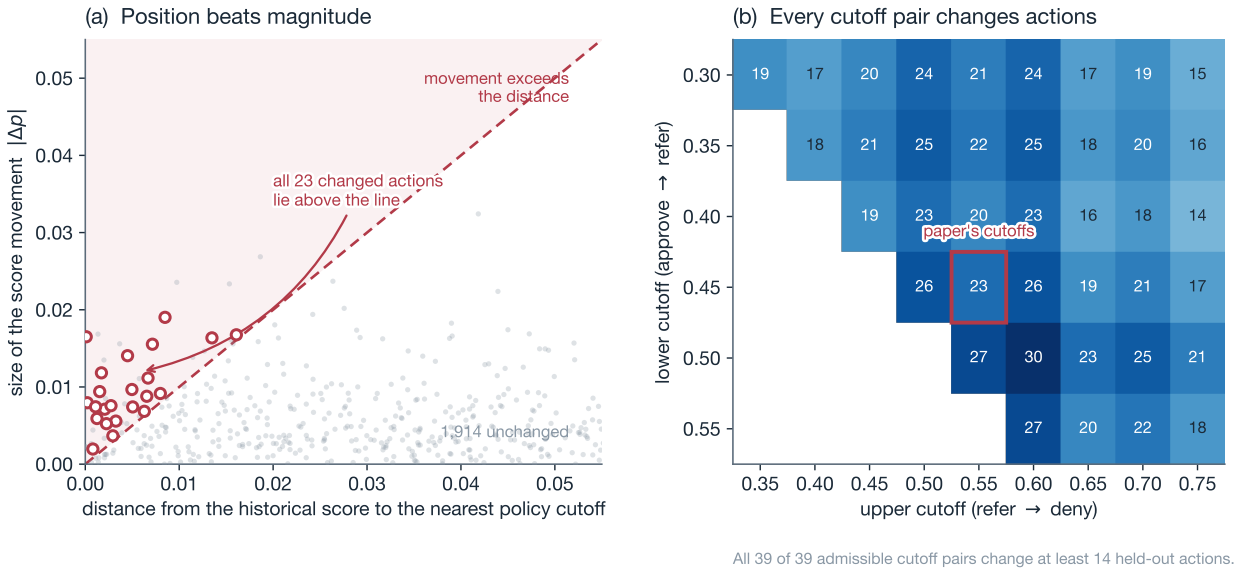


Figure 11: What makes a financial action change under a model refresh. **(a)** Each held-out application is plotted by its distance to the nearest policy cutoff under v_1 (horizontal) against the size of its score movement under v_2 (vertical). The 23 applications whose action changes (circles) all lie above the diagonal; the 1,914 unchanged applications (gray) are concentrated below it. The changed cases are distinguished by proximity to a cutoff, not by unusually large score movements. **(b)** Number of held-out actions that change, for every admissible pair of policy cutoffs on a .05 grid. All 39 pairs produce at least 14 changed actions. The pair used in this study (outlined) produces 23, close to the grid median of 21.

Faithfulness requires the historical executable. For the historical v_1 denial, three factors make the largest adverse local contributions. They are the percentage of trades never delinquent, the number of satisfactory trades, and the external risk estimate. We intervene on the preserved v_1 executable while holding all other inputs fixed. Improving each of these three factors independently lowers p below .55 and shifts the action from *deny* to *refer*, exactly as the stated reasons predict.

These probes demonstrate intervention-based faithfulness for the v_1 decision. The stated reasons are more than plausible descriptions of the denial. When each reported factor changes in the favorable direction, the preserved executable's score and action also change in the predicted direction. Their evidentiary value is therefore version-specific.

A reason statement reconstructed from v_2 could be fully faithful to the current referral and still fail to substantiate the historical denial. It would describe the decision function currently available, not the function that actually exercised authority at t_0. Historical explainability is therefore indexed to the historical executable. It cannot be established merely by generating a plausible explanation from the current model.

A foreseeable audit exposes the gap. Assume a foreseeable consumer-credit review in which the institution must substantiate the principal reasons given for the historical adverse action. If the institution retains only the final denial notice without preserving v_1 and its decision-time execution state, a current replay using v_2 will return *refer*. A rationale reconstructed from v_2 may accurately

describe the current model, but it is evidentially nonresponsive to the historical denial. Historical reproducibility fails, and the faithfulness of the historical explanation can no longer be validated.

By contrast, the preserved v_1 evidence bundle makes historical verification possible. It includes the decision-time inputs, fitted preprocessing state, model parameters, policy thresholds, and version-specific executable. With this bundle, the institution can reproduce the historical score and action and perform the faithfulness probes above. The experiment therefore identifies both sides of the mechanism: current-only retention opens the Verifiability Gap, whereas preservation of the historical executable restores the verification capacity needed to close it. Table 8 summarizes the evidence tests, and Figure 12 depicts the resulting decision-level gap.

Table 8: Credit-decision experiment: aggregate stability can conceal decision-level historical-fidelity failure.

Evidence test	**Observed result**	**Governance implication**
Version-level validation	$\text{AUC}_{v_1} = .7901$ and $\text{AUC}_{v_2} = .7899$; 23 of 1,937 held-out actions change, yielding $R_H = .9881$.	Stable aggregate discrimination does not guarantee historical action fidelity.
Case-level action	v_1: $p = .5580$, deny; v_2: $p = .5488$, refer. A score movement of .0092 crosses the fixed denial threshold.	Small numerical changes can produce discrete and consequential financial action changes at policy boundaries.
Current reproducibility	Each frozen version returns its own score and action in 10/10 executions.	$R_O = 1$ can coexist with case-level $R_H = 0$ because current reproducibility and historical fidelity evaluate different versions.
Faithfulness probes	Three one-factor interventions on v_1 lower the score below .55 and change deny to refer exactly as the stated reasons predict.	Reasons have evidentiary value only when tested against the executable that produced the historical action.
Retention condition	Current-only replay returns the v_2 referral; replay of the preserved v_1 evidence bundle recovers the historical denial and permits faithfulness testing.	Current-only retention opens the Verifiability Gap; preserving the historical executable restores verification capacity.

This experiment adds a critical controlled decision-level result to the multi-agent findings. Neither stochastic decoding nor inter-agent propagation is necessary for historical verification to fail. A deterministic and directly interpretable credit model can remain perfectly reproducible within each frozen version while a routine model refresh changes the financial action taken on the same case. At the decision level, R_O can equal one while R_H equals zero, even when aggregate predictive performance remains virtually unchanged. The relevant audit object is therefore not merely the current model. It is the specific historical executable against which the historical action and its stated reasons can still be tested.

8. Discussion

8.1 Theoretical Contributions

This paper makes three contributions to the IS literature on governing autonomous systems.

Verifiability as authority–evidence alignment. First, we bring explainability and reproducibility into a single governance construct. Prior IS research has primarily developed the explanatory route through inscrutability, sociotechnical envelopment, and blackboxing [Berente et al., 2021, Asatiani et al., 2021, Kronblad et al., 2024]. That work asks whether a responsible party can understand and

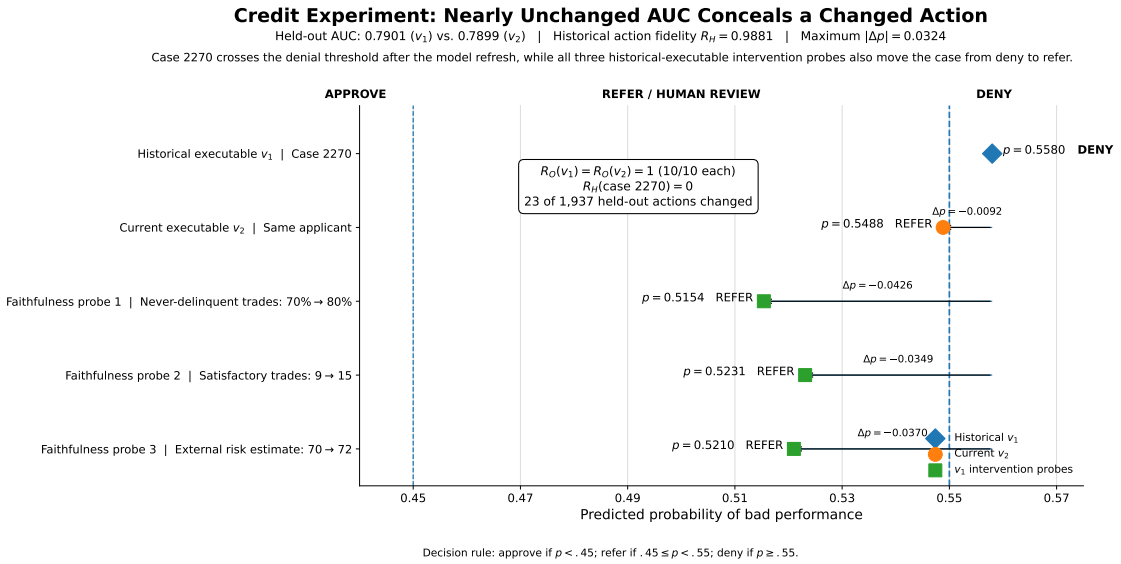


Figure 12: A decision-level Verifiability Gap under current-only retention. Version 1 returns $p = .5580$ and denies the application; version 2 returns $p = .5488$ and refers the same case. The small score movement crosses the fixed denial threshold. Both frozen versions have perfect current outcome reproducibility, yet repeated execution of the current version cannot recover the historical action. Preserving the version 1 executable, decision-time inputs, execution state, policy rule, and faithfulness probes restores the capacity to reproduce and substantiate the historical denial.

justify why a system acted. We identify a distinct empirical route: whether the relevant decision and its materially significant process can be reconstructed.

In conventional model governance, weakness in one route could often be partly offset by the other. Agentic systems weaken both because the operative decision object is distributed across models, tools, memory states, orchestration, inter-agent handoffs, and provider-controlled infrastructure. The *Verifiability Gap* captures the resulting shortfall between the evidence required by delegated authority and the verification capacity retained after the decision.

Verifiability is therefore not a static property of a model. It is a relation between a decision, a verifier, an evidentiary standard, and an audit lag, indexed by $q = (v, \sigma, \tau)$. The same decision may be readily verifiable to an internal engineer immediately after execution but no longer verifiable to an external regulator eighteen months later. This shifts the unit of analysis from model transparency to whether the complete decision system preserves sufficient evidence for the authority it exercises.

Reproducibility as a governance profile, not a scalar. Second, we separate the four reproducibility targets rather than reporting one number, and use verdict differentiation to diagnose output degeneracy:

$$\mathcal{R}(t;\sigma) = \big(R_O(t), R_H(t), R_P(t;\sigma), R_T(t); D_V(t)\big). \tag{24}$$

This profile exposes three forms of false assurance hidden within standard output-agreement metrics. A current decision may repeat even though the historical action is no longer recoverable from the current system alone. A final verdict may recur even when its exact recorded execution path is nonidentical. Exact nonidentity does not by itself establish material nonequivalence. It nevertheless prevents verdict agreement from serving as proof of process reproduction. Most perversely, outcome reproducibility may improve because the system collapses toward blanket actions

and ceases to distinguish among cases.

These findings revise the expected mechanism of serial agentic failure. Additional calls do not necessarily produce smooth, independent multiplicative decay. Serial pipelines can instead map nonidentical upstream states onto the same categorical action, preserving verdict agreement while leaving process equivalence unestablished. Study 3 extends this distinction to explainability: an intelligible rationale becomes historical verification evidence only when it behaviorally tracks the executable that produced the action under review.

Evidence-contingent delegation. Third, we specify when delegated authority remains defensible. Baird and Maruping [2021] theorize what delegation to an agentic artifact entails. The Verifiability Gap specifies what must remain true after delegation: an institution should authorize no more consequential action than it can later substantiate.

Evidence-contingent delegation therefore treats autonomy not as a one-time approval but as a revocable right tied to retained verification capacity. Provider releases, data refreshes, tool changes, control-surface changes, and orchestration updates are not merely technical maintenance when they alter either financial actions or the evidence available to defend them. They are governance events that may require the institution to strengthen evidence retention, reduce delegated authority, or both.

8.2 What the Three Studies Establish

Study 1: The institution may control neither the object nor the means of audit. Study 1 shows that provider-side change can affect both the historical decision object and the controls available for replay (§7.3). Across four dated releases, identical cases and instructions did not always recover the baseline modal action. During the same release line, the provider also withdrew a temperature control relevant to replay. The analytically separate local-control arm shows that the availability of temperature and seed controls can materially affect current outcome reproducibility.

The two arms do not constitute a single causal estimate. Together, however, they establish *control-surface dependence*: retained verification capacity can depend on execution controls that the deploying institution neither owns nor can unilaterally preserve. A firm may therefore lose both the historical decision object and part of the means needed to reconstruct it even when the firm itself changes nothing.

Study 2: Orchestration is a latent policy layer. Study 2 holds the locally served model, cases, and execution conditions fixed while changing architecture (§7.4). The one- and ten-agent configurations each achieve perfect current outcome reproducibility under temperature zero, yet their modal actions match in only 16 of 32 cases. Agent roles, topology, aggregation, and the information supplied to the final decider form a latent policy layer. Technical reconfiguration can therefore alter the operative financial decision rule without any explicit change to the stated policy.

The scale sweep identifies a second mechanism. The initial move from one to multiple agents reduces R_O, but its later rebound coincides with a collapse in verdict differentiation. At forty and fifty agents, $D_V = 0.086$, as the fixed concern-oriented roster increasingly maps different cases and nonidentical execution records onto a narrow set of conservative actions. Across the seven architectures, no case achieves exact full-record identity across all repetitions.

This is *degenerate outcome reproducibility*: apparent outcome stability improves as the system loses the discriminatory capacity it was authorized to exercise. Exact record nonidentity does not establish that every textual difference is materially consequential. It does establish that verdict agreement alone cannot demonstrate process reproduction.

Study 3: Perfect current replay does not recover history. Study 3 removes generative sampling and multi-agent orchestration as necessary explanations (§7.5). The two frozen logistic credit models share the same architecture, variables, estimation code, and policy thresholds. Their held-out discrimination is nearly identical: $\text{AUC}_{v_1} = .7901$ and $\text{AUC}_{v_2} = .7899$. Each model also perfectly reproduces its own score and action across repeated executions, yielding $R_O(v_1) = R_O(v_2) = 1$.

Nevertheless, 23 of 1,937 held-out actions change after the model refresh. For case 2270, v_1 returns $p = .5580$ and denies the application, whereas v_2 returns $p = .5488$ and refers it. The small score movement crosses the fixed policy threshold, producing case-level historical fidelity of $R_H = 0$. Current reproducibility is therefore a within-version property; historical fidelity is a cross-version property. Repeated execution of the current version can perfectly reproduce the current referral while never recovering the historical denial.

The intervention probes provide the complementary explainability result. The three principal adverse reasons behaviorally track the preserved v_1 executable: moving each factor in the favorable direction changes the historical score and action as predicted. Their evidentiary value is therefore version-specific. A rationale generated from v_2 may faithfully explain the current referral, but it cannot substantiate the historical v_1 denial.

Taken together, the controlled studies establish a stronger claim than "more agents are less stable." Current outcome agreement does not establish historical or process verification. An intelligible explanation also does not necessarily provide faithful historical evidence. Model-level and output-level stability metrics can therefore overstate the verification capacity retained for delegated authority. The studies identify these mechanisms; they do not estimate how frequently the mechanisms occur across deployed financial institutions.

8.3 Implications for Agentic AI Governance

Material change should trigger evidence-based reauthorization. A material change should be defined by its effects on financial behavior or verification capacity, not by whether the institution initiated it. Changes to models, data, prompts, tools, memory, execution controls, role rosters, topologies, aggregation rules, or policy thresholds can alter what the system does or whether prior decisions remain defensible.

After a material change, institutions should replay affected historical cases. They should assess $\mathcal{R}(t; \sigma)$ by decision family and inspect D_V for output collapse. Autonomous action should be reauthorized only where the retained evidence continues to support it. When verification capacity falls below the standard required by delegated authority, the institution must respond. It can strengthen the evidence bundle, reduce autonomous authority, or route affected decisions to substantive human review.

The relevant record is an executable evidence bundle. Static logs may document what occurred without preserving the capacity to reproduce or substantiate it. A defensible evidence bundle connects all material elements of a decision. The bundle includes decision-time inputs, fitted preprocessing state, model and component versions, prompts, policy thresholds, tool responses, retrieved documents, memory states, causally material handoffs, and the final action. Standardized documentation artifacts such as model cards and datasheets establish part of this record, but they describe a component rather than substantiate a particular decision [Mitchell et al., 2019, Gebru et al., 2018].

Historical reconstruction may also require *recorded-environment replay.* The system must be evaluated against the tool responses, retrieved documents, and relevant state available when the decision was made. The current environment is not necessarily an adequate substitute. Providers may not retain historical endpoints. Institutions should then seek contractual version retention,

model pinning, escrow, or functionally equivalent replay arrangements. These arrangements should follow foreseeable audit horizons rather than product release cycles.

Explainability controls must attach to the same evidence bundle. Research shows that fluent reasoning can be unfaithful [Turpin et al., 2023, Lanham et al., 2023]. Study 3 provides a tractable control prototype: intervene on the factors identified as reasons and test whether the preserved executable responds as the explanation predicts. For agentic chains, the same logic extends to materially relevant intermediate states, retrieved evidence, handoffs, and control decisions.

Provenance must compose across boundaries. Per-agent logs do not automatically constitute an end-to-end evidentiary record. Each causally material handoff should retain a persistent decision identifier, the sender and recipient, the transferred state, the applicable component version, and the transformation performed. Existing provenance research records the lineage of digital artifacts [Singh et al., 2019], while authenticated-delegation schemes can establish agent identity [South et al., 2025]. Agentic governance additionally requires these records to compose across models, agents, providers, and firms.

The requirement is not perfect textual reconstruction of every word. It is an evidence chain sufficient to establish the materially relevant decision process under the applicable verifier and evidentiary standard.

Missing evidence creates exposure independent of decision quality. A financial decision may be accurate and still be difficult to defend when the institution cannot reconstruct the evidence relevant to its production. Failure to preserve a reconstructable record can therefore create compliance and model-risk exposure independently of predictive accuracy [Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011].

Missing historical replay can also prevent a firm from bounding remediation. If a fault is discovered, the institution may be unable to determine whether the fault affected one decision or an entire decision population. Dependence on a shared provider adds a network-level exposure: a common provider change may create a correlated auditability shock across institutions even when their positions and customers differ [Financial Stability Board, 2024, European Central Bank, 2024]. Verifiability therefore affects not only retrospective explanation but also remediation cost, third-party concentration risk, and the defensibility of continuing autonomous operation.

8.4 Limitations

These studies are controlled mechanism tests rather than sector-wide prevalence estimates.

Study 1 covers one provider release line (Claude Opus), 32 constructed cases, four decision families, four dated releases, and a 185-day window. The release-timeline arm cannot separate behavioral drift from simultaneous capability change, identify silent change within a nominally unchanged version, or determine whether the same pattern generalizes across providers. Historical fidelity does not decline to one-half during the observation window, so an auditability half-life cannot yet be estimated.

Study 2 uses two locally served small models and one commercial frontier model whose parameter count the provider does not disclose, so model scale cannot be treated as a continuous covariate and is reported as a tier. It uses a researcher-designed serial architecture, a fixed concern-oriented role roster, frozen tool states, and a single-pass graph. It therefore isolates specialization, causal state transfer, decoding controls, and orchestration, but not live retrieval, persistent memory, autonomous goal decomposition, iterative planning, or self-correction. Its categorical actions may also behave differently from continuous financial decisions such as prices, limits, or position sizes. Exact record identity is an intentionally stringent diagnostic of textual reproduction; it does not by itself establish

material process nonequivalence.

Study 3 uses one public dataset, five selected variables, one transparent logistic model class, one staged refresh, and illustrative policy thresholds. It is a mechanism test of historical replay and explanation faithfulness, not a production underwriting policy, fair-lending validation, or estimate of lender behavior. The detailed intervention analysis centers on a selected denial-to-referral boundary case. Replication should examine the full population of changed actions and additional datasets, model classes, policy rules, and refresh events.

The studies also contain no organizational deployment data. They directly identify the material-change mechanism relevant to P4 and provide a bounded operationalization of historical explanation faithfulness relevant to P5. Study 2 provides a system-internal analogue of the serial evidentiary dependence in P6, while Study 1 identifies a provider-side antecedent relevant to P7. These studies do not directly observe several proposed organizational and network mechanisms. They do not show how firms set autonomy ceilings in P1, compare fault-escape rates across controls in P2, or estimate the verifiability–efficiency and deskilling effects in P3. Nor do they track evidence custody across firms in P6 or demonstrate synchronized institutional failure in P7. The empirical findings should therefore be interpreted as mechanism evidence, not as estimates of organizational or network outcomes.

8.5 Future Agentic Governance Research

Measure longitudinal verifiability decay. Future studies should estimate how verification capacity changes over audit lags across multiple providers, model families, decision types, and control surfaces. They should preserve raw repetitions, record observed material-change events, probe nominally unchanged versions for silent drift, and stratify cases by their distance from policy boundaries. Such designs could determine whether historical fidelity decays gradually or through discrete losses of models, controls, data states, and provider support, and could yield context-specific estimates of auditability half-life.

Separate orchestration from world coupling. Future experiments should vary agent count, causal depth, topology, role composition, aggregation rules, and decider information independently. They should then vary live retrieval, tool responses, persistent memory, iterative planning, and self-correction while retaining the historical environment for replay. Such designs would distinguish instability arising from generation, causal state transfer, world coupling, and categorical compression. They should also extend the analysis from categorical verdicts to continuous financial actions and develop preregistered standards for materially equivalent, rather than merely textually identical, execution paths.

Connect explanation faithfulness to human governance. Study 3 provides a tractable intervention-based operationalization of historical faithfulness in a transparent model. The next step is to test explanations generated by agentic systems against the historical chains that produced their actions. Behavioral experiments should compare faithful explanations, fluent but unfaithful rationales, action bounds, decision telemetry, and combinations of these controls.

Relevant outcomes include reviewer fault detection, fault escape, confidence calibration, challenge behavior, review latency, and reliance on automated recommendations. Longitudinal studies should further examine whether replay and intervention practices preserve reviewer expertise or whether dependence on agent-generated explanations gradually produces deskilled oversight.

Study organizational and network governance. Field research should examine how model-risk, compliance, legal, and business functions set autonomy ceilings. It should also identify when a change

to a model, tool, data source, or architecture requires reauthorization. Such work would directly test P1 and permit estimation of the verifiability–efficiency and deskilling mechanisms in P3.

At the network level, research should trace evidence custody across interorganizational agent handoffs. It should also test whether provenance standards, contractual replay rights, version retention, escrow, and independent fallbacks reduce correlated auditability shocks from shared providers. Quantifying the option value of historical replay would further connect the Verifiability Gap to remediation costs, insurance, capital allocation, contracting, and third-party concentration risk. Together, these studies would move agentic governance from static model approval toward a longitudinal theory of how authority, evidence, and interorganizational dependence must be governed together.

9. Conclusion

Agentic AI changes FinTech governance because AI systems no longer merely support financial judgment; they increasingly coordinate models, tools, data, and other agents to exercise delegated financial authority. The central governance challenge is therefore not simply whether a model is accurate or whether a human appears somewhere in the workflow. It is how institutions calibrate agent authority, preserve evidence of how that authority was exercised, and trigger substantive human intervention when retained verification capacity no longer supports autonomous action.

We theorize this problem through the *Verifiability Gap*: the positive shortfall between the verification demanded by delegated authority and the explainability and reproducibility retained after a decision. We operationalize the construct's central reproducibility mechanisms. We separate the four reproducibility targets rather than reporting one number. We also use verdict differentiation to diagnose apparent stability produced by output collapse. From this framework follows *evidence-contingent delegation*: agentic authority should remain valid only while the institution retains sufficient evidence to substantiate the decisions made under that authority.

The three studies identify complementary mechanisms through which that condition can fail. Study 1 shows that provider-side releases can alter historical financial actions while also changing the control surface available for replay. Study 2 shows that multi-agent orchestration operates as a latent policy layer. Changes in roles and topology can change financial decisions. Stable terminal actions can coexist with nonidentical execution records. Outcome reproducibility can also improve as the system collapses toward blanket defaults. Study 3 shows that even a deterministic and transparent credit model can perfectly reproduce its current action while failing to reproduce a historical action. Its intervention probes add the corresponding explainability result. A stated reason becomes historical verification evidence only when it can be tested against the preserved executable that produced the decision.

These findings clarify the role of human-in-the-loop governance. Human oversight is not substantive merely because a reviewer approves a final output. A reviewer may receive only a fluent rationale, a terminal verdict, or a replay from the current model. That evidence may not reveal that the historical action, decision path, or operative reasons have changed. Meaningful oversight requires an executable evidence bundle. It must connect decision-time inputs, historical component versions, relevant execution and memory states, tool responses, causally material handoffs, policy rules, and final actions. It must also support faithfulness tests that link stated reasons to the historical decision process.

Human intervention should therefore be evidence-contingent rather than uniformly imposed. Where retained verification capacity remains commensurate with the authority delegated, autonomous execution may continue. A material change may weaken verification capacity by altering the model, data, tools, control surface, memory, or orchestration. The governance response should

then include historical replay, renewed faithfulness testing, tighter action bounds, or substantive human approval. Where the institution can no longer reproduce or explain a decision to the required standard, delegated authority should contract rather than persist by default.

Our controlled studies identify these mechanisms rather than estimate their prevalence across deployed financial institutions. Future agentic governance research should address four questions. How do firms set and revise autonomy thresholds? Which human-review designs reduce fault escape rather than produce ceremonial approval? Does agent-assisted oversight preserve or erode human expertise? How can executable evidence and replay rights be maintained across organizational and provider boundaries?

FinTech agentic AI governance is the dynamic allocation of authority between agents and humans according to retained verification capacity. Its goal is neither maximum autonomy nor a human checkpoint attached to every action. It is *defensible autonomy*. Agents act where their exercise of authority remains verifiable. Humans intervene where verification capacity falls below the required standard. Institutions preserve sufficient evidence to substantiate what was done, through what process, for what reasons, and on what basis.

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interest. The author declares no competing financial or personal interests.

Data availability. The code and data reproducing every empirical figure and table are provided in the accompanying `pub-code` package and are available at `https://anonymous.4open.science/status/AgenticAI-AEAF`. The stored records regenerate every reported statistic offline, without an API key. Re-executing the hosted arms is a different matter and will not return the same values, because the endpoints are versioned by the provider and `claude-opus-5` is an undated alias; that limitation is itself one of the findings. The package contains four groups of records. The single-call waves (`wave0_*.jsonl`) contain 1,600 runs across a local model, a smaller hosted model, and `claude-opus-5`. The cross-version study (`versions_runs.jsonl`) contains 128 runs and uses the modal verdict over three repetitions per release. The agent experiments are stored in five files. `agents_runs.jsonl` contains 640 runs comparing one and ten agents. `agents_temp0_runs.jsonl` contains 160 runs that isolate the random seed from temperature. `agents_sweep.jsonl` contains 1,120 runs across seven agent counts, and `llama8b/` contains the 2,560-run replication of those arms at the larger local scale. `agents_sonnet5.jsonl` contains the 288 runs of the frontier comparison. A further file, `agents_opus5_pilot.jsonl`, holds a 64-run cost-calibration pilot at one repetition; it supports no reported statistic and is included only for completeness. Every agent's output is stored for every run, so the record-reproduction results can be recomputed rather than taken on trust. The 32 decision cases are constructed rather than drawn from live portfolios and are reconstructible from code (case-set hash `696afd10`); the agent configuration is documented in the Supplemental Material appended to this article (Section S1).

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. arXiv:1810.03292v3, 2018.

Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. Contestable AI by design: Towards a framework. *Minds and Machines*, 33(4):613–639, 2023.

Aleksandre Asatiani, Pekka Malo, Per Rådberg Nagbøl, Esko Penttinen, Tapani Rinta-Kahila, and Antti Salovaara. Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems*, 22(2):325–352, 2021. doi: 10.17705/1jais.00664.

Berk Atıl, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. Non-determinism of "deterministic" LLM system settings in hosted environments. In *Proceedings of the Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)*, pages 135–148, 2025. arXiv:2408.04667.

Aaron Baird and Likoebe M. Maruping. The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly*, 45(1):315–341, 2021. doi: 10.25300/MISQ/2021/15882.

Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics*, 143(1):30–56, 2022. doi: 10.1016/j.jfineco.2021.05.047.

Nicholas Berente, Bin Gu, Jan Recker, and Radhika Santhanam. Managing artificial intelligence. *MIS Quarterly*, 45(3):1433–1450, 2021.

Tobias Berg, Valentin Burg, Ana Gombović, and Manju Puri. On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *The Review of Financial Studies*, 33(7):2845–2897, 2020. doi: 10.1093/rfs/hhz099.

Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. Supervisory guidance on model risk management (SR 11-7). Technical report, Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou,

Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258v3, 2021.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for Variance in Machine Learning Benchmarks. arXiv:2103.03098v1, 2021.

Laura B. Cardinal, Sim B. Sitkin, and Chris P. Long. Balancing and Rebalancing in the Creation and Evolution of Organizational Control. *Organization Science*, 15(4):411–431, 2004. doi: 10.1287/orsc.1040.0084.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness Accountability and Transparency*, pages 651–666, 2023. doi: 10.1145/3593013.3594033.

Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. arXiv:2102.04201v2, 2021.

Nicholas Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016. doi: 10.1145/2844110.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608v2, 2017.

European Central Bank. The rise of artificial intelligence: Benefits and risks for financial stability. Technical report, Financial Stability Review, Special Feature, May 2024.

European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L series, 12 July 2024.

Samer Faraj, Stella Pachidi, and Karla Sayegh. Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1):62–70, 2018. doi: 10.1016/j.infoandorg.2018.02.005.

FICO. Explainable machine learning challenge: Home equity line of credit (HELOC) dataset. Dataset and challenge documentation, 2018.

Financial Industry Regulatory Authority. Emerging trend in GenAI: Observations on AI agents. FINRA, January 27, 2026.

Financial Industry Regulatory Authority. Regulatory Notice 24-09: FINRA reminds members of regulatory obligations when using generative artificial intelligence and large language models. Technical report, FINRA, 2024.

Financial Stability Board. The financial stability implications of artificial intelligence. Technical report, Financial Stability Board, November 2024.

Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quarterly*, 45(3):1527–1556, 2021.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *Journal of Finance*, 77(1):5–47, 2022.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. Datasheets for Datasets. arXiv:1803.09010v8, 2018.

Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.

Itay Goldstein, Wei Jiang, and G Andrew Karolyi. To FinTech and Beyond. *The Review of Financial Studies*, 32(5):1647–1661, 2019. doi: 10.1093/rfs/hhz025.

Ben Green and Yiling Chen. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019. doi: 10.1145/3359152.

Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 23(4):497–530, 1999.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11503.

H. Han. Challenges of reproducible AI in biomedical data science. *BMC Medical Genomics*, 18 (Suppl 1):8, 2025.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4198–4205, 2020.

JPMorganChase. Letter to shareholders from Mary Callahan Erdoes: Asset and Wealth Management. In *2025 Annual Report*, JPMorgan Chase & Co., 2026.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Katherine C. Kellogg, Melissa A. Valentine, and Angéle Christin. Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, 14(1):366–410, 2020. doi: 10.5465/annals.2018.0174.

Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010. doi: 10.1016/j.jbankfin.2010.06.001.

Laurie S. Kirsch. Portfolios of Control Modes and IS Project Management. *Information Systems Research*, 8(3):215–239, 1997. doi: 10.1287/isre.8.3.215.

Charlotta Kronblad, Anna Essén, and Magnus Mähring. When justice is blind to algorithms: Multilayered blackboxing of algorithmic decision-making in the public sector. *MIS Quarterly*, 48 (4):1637–1662, 2024.

Naveen Kumar, Xiahua Wei, and Han Zhang. Agentic artificial intelligence as a new frontier in information systems: Promise, peril, and research opportunities. *Information & Management*, 63 (3):104317, 2026. doi: 10.1016/j.im.2026.104317.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Zachary C. Lipton. The Mythos of Model Interpretability. *Queue*, 16(3):31–57, 2018. doi: 10.1145/3236386.3241340.

Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151: 90–103, 2019. doi: 10.1016/j.obhdp.2018.12.005.

Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874v2, 2017.

Mastercard. Mastercard unveils Agent Pay, pioneering agentic payments technology to power commerce in the age of AI. Company announcement, April 29, 2025.

Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183, 2004.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019. doi: 10.1145/3287560.3287596.

Alex Murray, Jen Rhymer, and David G. Sirmon. Humans and Technology: Forms of Conjoined Agency in Organizations. *Academy of Management Review*, 46(3):552–571, 2021. doi: 10.5465/amr.2019.0186.

William G. Ouchi. Markets, Bureaucracies, and Clans. *Administrative Science Quarterly*, 25(1): 129–141, 1980. doi: 10.2307/2392231.

Raja Parasuraman and Victor Riley. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2):230–253, 1997. doi: 10.1518/001872097778543886.

Roger D. Peng. Reproducible Research in Computational Science. *Science*, 334(6060):1226–1227, 2011. doi: 10.1126/science.1213847.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Edward Raff. A Step Toward Quantifying Independently Reproducible Machine Learning Research. arXiv:1909.06674v1, 2019.

Arun Rai, Panos Constantinides, and Suprateek Sarker. Editor's comments: Next-generation digital platforms: Toward human–AI hybrids. *MIS Quarterly*, 43(1):iii–ix, 2019.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. arXiv:2001.00973v1, 2020.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, 2016. doi: 10.18653/v1/n16-3020.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Jatinder Singh, Jennifer Cobbe, and Chris Norval. Decision provenance: Harnessing data flow for accountable systems. *IEEE Access*, 7:6562–6574, 2019.

Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999. doi: 10.1006/ijhc.1999.0252.

Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. Authenticated delegation and authorized AI agents. *arXiv preprint arXiv:2501.09674*, 2025.

Mike H. M. Teodorescu, Lily Morse, Yazeed Awwad, and Gerald C. Kane. Failures of Fairness in Automation Require a Deeper Understanding of Human–ML Augmentation. *MIS Quarterly*, 45 (3):1483–1500, 2021. doi: 10.25300/misq/2021/16535.

Haileleol Tibebu. The accountability horizon: An impossibility theorem for governing human-agent collectives. *arXiv preprint arXiv:2604.07778*, 2026.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Visa. Find and buy with AI: Visa unveils new era of commerce. Company announcement, April 30, 2025.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. In *SSRN Electronic Journal*, 2017. doi: 10.2139/ssrn.3063289.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A Survey on Large Language Model based Autonomous Agents. arXiv:2308.11432v7, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35*, pages 24824–24837, 2022. doi: 10.52202/068431-1800.

Martin Wiener, Magnus Mähring, Ulrich Remus, and Carol Saunders. Control Configuration and Control Enactment in Information Systems Projects: Review and Expanded Theoretical Framework. *MIS Quarterly*, 40(3):741–774, 2016. doi: 10.25300/misq/2016/40.3.11.

Maranke Wieringa. What to account for when accounting for algorithms. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 1–18, 2020. doi: 10.1145/3351095.3372833.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. TradingAgents: Multi-agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629v3, 2022.

Supplemental Material

This appendix provides the detailed configuration for the controlled multi-agent experiment reported in Study 2.

S1. The Multi-Agent Experiment

This supplement documents the multi-agent configuration used in Study 2. Everything needed to rebuild and re-run it is provided here. The replication package contains the implementation file `run_agents.py`.

S1.1 What Has to Be Held Fixed, and What Must Not Be

A reproduction measurement is only meaningful if the thing being varied is the thing an auditor cannot control. Two decoding settings govern this.

The first is the random seed. When a provider accepts a fixed seed, sampling can become reproducible under a fixed executable. A fixed-seed measurement therefore reports implementation determinism. It does not reproduce the conditions faced by a later verifier. In deployment, the decision-time seed may not be retained or exposed. Our main comparisons therefore vary the seed across repetitions.

The second is temperature. Pinning temperature to zero is the strongest available control over how the model chooses its words. It is also the control that the provider withdrew in the release line examined in Study 1. When it is gone, what an auditor gets is the provider's default. We therefore report the condition in which temperature is not sent at all.

One further detail matters for reading conditions B and C. Within a single decision, every call in the chain receives the same seed. These conditions therefore understate divergence relative to a deployment in which each call is sampled independently. Condition D removes the seed entirely and is the least controlled condition. The four conditions are summarized in Table S1:

Table S1: Decoding conditions. Condition A is an implementation check; Conditions B–D progressively remove controls relevant to later replay.

	Temperature	**Seed**	**What it measures**
A	pinned to 0	fixed	Whether the implementation is deterministic.
B	pinned to 0	varied	Whether the seed matters once temperature is pinned.
C	not sent	varied	The effect of releasing the temperature alone.
D	not sent	not sent	What an auditor faces: neither control is available.

S1.2 The Multi-Agent Configuration

The orchestration arm implements the governance logic shown in Figure 5. Direct access to the original case is restricted to the first layer. Upstream roles are report-only, and every designated handoff is retained. Only the terminal Decider may issue the financial action. Scaled configurations vary the number of specialist calls while preserving these restrictions. Table S2 reports the canonical role functions. The replication package generates the exact graph, role multiplicities, and call count used at each scale. Its configuration files should be treated as the authoritative specification.

Each model call receives only the inputs designated by the graph. The terminal action is therefore produced from retained upstream reports rather than from an unrecorded vote or direct access to the original case. We record every role output and handoff. This makes it possible to identify the first exact divergence. It also shows whether later stages absorb, propagate, or compress that divergence before the final verdict.

Table S2: Canonical role functions in the controlled multi-agent configuration.

Agent	**Layer**	**Reads**	**Instruction**
retrieval	1	the case	Extract every decision-relevant fact and number. Do not judge.
market	1	the case	Summarize the market and environment signals. Do not judge.
policy	1	the case	State the policy thresholds and rules that apply. Do not judge.
analyst	2	retrieval, policy	Assess quantitatively against the stated thresholds. Do not choose an action.
risk	2	retrieval, market	State the material risks and an overall severity. Do not choose an action.
compliance	2	retrieval, policy	Flag any regulatory or fair-treatment concern. Do not choose an action.
aml	2	retrieval	Flag any financial-crime or counterparty concern. Do not choose an action.
quant	3	analyst, risk	Score each allowed option on expected benefit and downside. Do not choose.
customer	3	compliance, aml	State the effect on the customer or counterparty. Do not choose.
critic	4	quant, customer	Challenge the emerging view; name the strongest argument against it. Do not choose.
decider	5	quant, customer, critic	Choose exactly one allowed label and give a one-sentence reason.

S1.3 Execution

Both agent configurations used the same locally served model (`llama3.2:3b`) on the same machine. They also used the same 32 fixed cases: eight each in credit, trade, AML, and rebalancing. A locally served model is the right control: it removes the provider from the picture entirely, so that whatever divergence appears cannot be attributed to an upstream update.

Exact call counts for each scale are generated from the replication configuration files rather than entered manually in the manuscript.

Alongside the verdict, we record every agent's output for every run. This design supports a second measurement that a single call cannot yield. We identify the earliest agent whose text differs across repetitions of the same case. That point locates where irreproducibility enters and how far it travels before reaching the verdict.

S1.4 What This Configuration Does and Does Not Implement

This arm isolates orchestration. It fixes the communication graph within each configuration, freezes tool payloads, and traverses the graph once. It excludes live retrieval, persistent memory, autonomous goal decomposition, iteration, and self-correction. Provider-side and live-environment

changes are examined separately in Study 1. The orchestration results should therefore be interpreted as a bounded estimate of architecture-induced divergence with the external environment held fixed.

S2. Proofs and Formal Calibration

This appendix proves the results stated in Sections 4.4 and 7.1, records one further consequence, and calibrates the quantities against the experimental data.

S2.1 Notation

Calls are indexed $i = 1, \dots, m$ in a topological order of the communication graph; S_k denotes the outputs of layer k, T the complete trace, and $Y = g(T)$ the action, recovered by a fixed parser from the terminal Decider output. For two independent repetitions of the same case, $R_T = \Pr[T^{(1)} = T^{(2)}]$ and $R_O = \Pr[Y^{(1)} = Y^{(2)}]$. Define the conditional self-agreement of call i,

$$p_i = \Pr\big[Z_i^{(1)} = Z_i^{(2)} \;\big|\; Z_j^{(1)} = Z_j^{(2)} \;\forall j < i\big], \tag{S1}$$

and the compression probability $\kappa = \Pr[Y^{(1)} = Y^{(2)} \mid T^{(1)} \neq T^{(2)}]$.

Lemma 1. *(i) $R_T = \prod_{i=1}^{m} p_i$, so R_T is non-increasing in m and $R_T \le \bar{p}^{\,m}$ whenever $p_i \le \bar{p}$. (ii) $R_O = R_T + \kappa(1 - R_T) \ge R_T$; in particular $R_O = \kappa$ when $R_T = 0$. (iii) The estimator used in the main text satisfies $\widehat{R}_O = |\mathcal{C}|^{-1} \sum_c \exp(-H_\infty(\hat{q}_c))$, where q_c is the verdict distribution across repetitions of case c and $H_\infty(q) = -\log \max_\ell q_\ell$.*

Proof. *(i)* Let $A_i = \{Z_i^{(1)} = Z_i^{(2)}\}$. Exact trace identity is $\bigcap_{i \le m} A_i$, and the chain rule gives $\Pr[\bigcap_i A_i] = \prod_i \Pr[A_i \mid \bigcap_{j<i} A_j] = \prod_i p_i$ by (S1). The conditioning in (S1) is what makes this an identity rather than an independence assumption. *(ii)* Y is a deterministic function of T, so trace identity implies verdict identity; total probability over $\{T^{(1)} = T^{(2)}\}$ and its complement gives the decomposition. *(iii)* $\max_\ell q_\ell = \exp(-H_\infty(q))$, averaged over cases. □

S2.2 Proofs of the reversibility results

Proof of Theorem 1. The inequalities are the data-processing inequality applied along Equation 3, which is valid precisely because no layer beyond the first reads C; the final step adds the deterministic map $Y = g(T)$. Iterating the strong data-processing inequality, whose coefficient η_k is strictly below one whenever the layer-k channel maps distinct inputs to overlapping output distributions, yields the geometric bound. Corollary 1 is the case $k > 1$ of the same chain. □

Proof of Theorem 2. (i)⇔(iv): with E_k trivial, the map $S_k \mapsto [S_{k-1}]_\sigma$ of Definition 1 is well defined exactly when no two materially distinct upstream states share a downstream image. (iv)⇒(ii): an injective relabelling preserves mutual information. (ii)⇒(iii): equality in the data-processing inequality holds iff $C \to S_k \to S_{k-1}$ is also a Markov chain, which is the definition of sufficiency. (iii)⇒(iv) modulo $\equiv_\sigma$: a sufficient statistic induces no collapse of materially distinct states. □

Proof of Theorem 3. Non-injectivity modulo $\equiv_\sigma$ means there are $s \not\equiv_\sigma s'$ with the same image. The map of Definition 1 is then not well defined, and because materiality means C is not independent of the distinction between s and s', the inequality $I(C; S_k) < I(C; S_{k-1})$ is strict, i.e. $\eta_k < 1$. The converse is Theorem 2(i)⇔(iv). □

Proof of Theorem 4. Reversibility requires $[T]_\sigma$ to be a function of (Y, E) almost surely, so $H([T]_\sigma \mid Y, E) = 0$. Then

$$H([T]_\sigma \mid Y) \;\le\; H([T]_\sigma \mid Y) + H(E \mid Y, [T]_\sigma) \;=\; H(E \mid Y) + H([T]_\sigma \mid Y, E) \;=\; H(E \mid Y) \;\le\; H(E).$$

Achievability is the standard construction: a code for the conditional distribution of the pre-image within each stage attains the conditional entropy, and the stage-wise terms compose by the chain rule. Corollary 2 is the definition of G_{dq} with these two quantities substituted, and Corollary 3 follows because Theorem 3 makes each compressing stage contribute a strictly positive term. □

S2.3 Proofs of the metric results

Proof of Theorem 5. For any distribution on L labels, $\max_\ell q_\ell \ge 1/L$, so $\widehat{R}_O \ge 1/L$ by Lemma 1(*iii*); at the population level $\Pr[Y^{(1)} = Y^{(2)}] = \sum_\ell q_\ell^2 \ge 1/L$ by Cauchy–Schwarz. The two limits are Lemma 1(*i*) and Theorem 1. Corollary 4 follows by exhibiting the constant policy, for which $q_c = \delta_{\ell_0}$ for every c, giving $R_O = 1$ and $I(C; Y) = 0$. □

Proof of Theorem 6. Under $\mathcal{A}$, $q_c = \delta_{\ell^\star(c)}$; under $\mathcal{B}$, $q_c = \delta_{\ell_0}$. Both are point masses, so for each case the empirical distribution over K repetitions is a point mass almost surely, and any functional of $\{q_c\}$ that does not reference label identities across cases evaluates identically under the two systems. Statistics that do reference cross-case label identity are excluded by hypothesis, which is Corollary 5. □

S2.4 Non-monotonicity of outcome reproducibility

Theorem 7. *With $\Delta f(n) = f(n{+}1) - f(n)$,*

$$\Delta R_O(n) = \underbrace{(1 - \kappa_{n+1})\,\Delta R_T(n)}_{\le 0} + \underbrace{(1 - R_T(n))\,\Delta\kappa(n)}_{\textit{sign of } \Delta\kappa}. \tag{S2}$$

Whenever added depth coarsens the terminal map enough that the second term dominates, R_O rises while R_T falls. The independent-compounding benchmark is the restriction $\kappa_n \equiv \kappa$, which deletes the second term.

Proof. Differencing Lemma 1(*ii*) and collecting terms gives (S2); the sign of the first term follows from Lemma 1(*i*). □

Theorem 7 predicts both the shape of the agent-count sweep and the pairing that the data exhibit: local maxima of R_O should coincide with local minima of $I(C; Y)$, because both are driven by the same coarsening of g. It also identifies the assumption that the dashed benchmark in Figure 9 violates.

The measured data sit in a degenerate case of (S2) that is worth stating plainly. Since $R_T = 0$ at every depth and at both model scales, $\Delta R_T(n) = 0$, the first term vanishes, and $\Delta R_O(n) = \Delta\kappa(n)$ exactly. The trace-decay force is not competing with compression in the observed range; it reached its floor at one agent and has no further room. Everything the sweep shows about outcome reproducibility is therefore a statement about compression alone. The two-term decomposition is what tells us this, and a configuration with $R_T > 0$ — a shallower pipeline, a pinned execution environment — is where the first term would become observable.

S2.5 Empirical calibration

Table S3 reports pairwise estimates from the frozen 32-case test bed, $k = 5$ repetitions, seeds 1000–1004, with identical code and cases across model scales. Mutual information is computed within decision family under a uniform case prior.

Table S3: Pairwise trace identity, outcome agreement, implied compression, and discrimination. Released temperature is the condition an auditor faces once the provider withdraws the control. With $L = 2$ or 3 per family, the floor of Theorem 5 lies between .33 and .50.

Model	**Condition**	R_T	R_O	κ	$I(C;Y)$ **(bits)**
`llama3.2:3b`	released, $n = 1$	0.0000	0.8313	0.8313	0.2986
	released, $n = 5$	0.0000	0.5938	0.5938	0.2130
	released, $n = 10$	0.0000	0.6594	0.6594	0.2118
	released, $n = 20$	0.0000	0.7844	0.7844	0.1429
	released, $n = 30$	0.0000	0.6687	0.6687	0.1781
	released, $n = 40$	0.0000	0.7875	0.7875	0.0853
	released, $n = 50$	0.0000	0.7344	0.7344	0.1990
`llama3.2:3b`	temperature 0, $n = 1$	1.0000	1.0000	—	—
	temperature 0, $n = 10$	0.9375	1.0000	1.0000	—
`llama3.1:8b`	released, $n = 1$	0.0000	0.7875	0.7875	0.2434
	released, $n = 5$	0.0000	0.5813	0.5813	0.2474
	released, $n = 10$	0.0000	0.6281	0.6281	0.2591
	released, $n = 20$	0.0000	0.5375	0.5375	0.1076
	released, $n = 30$	0.0000	0.5281	0.5281	0.1366
	released, $n = 40$	0.0000	0.6562	0.6562	0.1494
	released, $n = 50$	0.0000	0.5312	0.5312	0.1105
`llama3.1:8b`	temperature 0, $n = 1$	0.9688	1.0000	1.0000	—
	temperature 0, $n = 10$	0.4500	0.9125	0.8409	—

The decoupling regime is the empirical regime. Under released temperature, $R_T = 0$ at every depth and both model scales. By Lemma 1(ii) the outcome column *is* the compression column. Verdict agreement in this regime measures how strongly the terminal map collapses distinct traces, and nothing else.

The predicted pairing appears. The largest sweep value of outcome reproducibility, $R_O = .7875$ at $n = 40$, coincides with the smallest discrimination, $I(C;Y) = .0853$ bits, under a third of the single-agent value. This is the pairing Theorem 7 predicts and the empirical signature of Corollary 4: the configuration scoring best on the audit metric retained the least information about which case it was deciding.

A floor on destroyed information. Under released temperature, every case produced five distinct traces in five repetitions, collapsing to 1.4–1.8 distinct verdicts. The terminal map therefore destroyed at least $\log_2(5/1.44) \approx 1.8$ bits per episode at 3B and $\log_2(5/1.81) \approx 1.5$ bits at 8B. These are floors implied by five samples of a vastly larger trace space, not estimates of $H([T]_\sigma \mid Y)$; the true requirement is larger. An institution retaining only the final action retains none of the destroyed information, so $G_{dq} > 0$ by Corollary 2 for every episode in the experiment.

Composition amplifies a per-call defect. For `llama3.1:8b` at temperature 0 the single-call estimate is $\hat{p} \approx .9688$. Constant p_i would predict $R_T \approx .9688^{11} \approx .70$ at eleven calls; the observed value is .45. Because Lemma 1(i) is an identity, the shortfall is not estimation error: p_i must fall for calls

conditioned on longer upstream context. This is a quantitative form of the serial-decay mechanism obtained without further assumptions.

Estimation caveat. With $K = 5$ repetitions and eight cases per family, plug-in mutual information is upward-biased and noisy; the $n = 50$ value should not be read as a reversal. The theory constrains the bound in Theorem 1, not the sample path, and the comparison that carries weight is between the extremes rather than between adjacent depths.

S2.6 What does not follow: model scale

Model scale enters this theory in two places only: the per-call parameters p_i of (S1), and the contraction coefficients η_k induced by how a given model summarizes. Neither is a monotone function of parameter count, and no such monotonicity is observed here. The defensible claim is narrower and concerns controls rather than capability: retained process reproducibility is governed by whether the deployed configuration achieves $p_i = 1$, not by how capable the model is. Because $R_T = \prod_i p_i$, any deployment with $p_i < 1$ loses exact process reproducibility geometrically in causal depth; and because discrimination contracts along the same chain, a more capable model composed more deeply is not thereby more auditable.

S3. Provider Concentration in U.S. Mortgage Underwriting

P7 assumes that many institutions depend on the same upstream provider. This section reports whether that structure is present in a market where the dependency is publicly recorded. It is descriptive. Nothing here identifies the effect of a provider change; it establishes only that the exposure P7 presupposes exists at scale.

Every U.S. mortgage application covered by the Home Mortgage Disclosure Act is filed with the applicant's outcome, the reporting institution, and—since the 2018 reporting revision—the automated underwriting system that evaluated it. That last field makes provider dependence directly observable rather than inferred. We use the public loan-level files for 2019–2023 in the six largest reporting states (CA, TX, NY, FL, IL, PA), restricted to applications that received a credit decision (originated, approved but not accepted, or denied): 24,439,380 applications filed by 4,487 institutions. The overall denial rate is 20.1%, and debt-to-income ratio is the most frequently reported primary denial reason (31.6%).

Table S4: Automated underwriting systems in U.S. mortgage applications, 2019–2023, six states. Shares are of applications evaluated by *some* automated system; applications reporting no such system are excluded from the share denominator. The Herfindahl–Hirschman index is computed over system shares.

Year	AUS-decided applications	Desktop Underwriter *(Fannie Mae)*	Loan Product Advisor *(Freddie Mac)*	TOTAL Scorecard *(FHA)*	Top two	HHI
2019	2,929,098	63.9%	15.9%	8.2%	79.8%	4,535
2020	4,818,652	64.8%	23.9%	5.2%	88.7%	4,832
2021	5,149,015	64.4%	25.6%	4.4%	90.0%	4,834
2022	2,085,854	61.8%	22.2%	8.6%	84.0%	4,416
2023	1,582,341	55.8%	21.9%	11.9%	77.7%	3,786

Two features of Table S4 bear on P7. Concentration is high and persistent: two systems, both

owned by government-sponsored enterprises, evaluate between 77.7% and 90.0% of automated decisions in every year, and the Herfindahl–Hirschman index stays between 3,786 and 4,834 throughout. Dependence is also broad rather than confined to large filers. In 2021, 1,674 distinct institutions routed applications through Desktop Underwriter and 1,241 through Loan Product Advisor.

Dependence is concentrated at the level of the individual institution as well as the market. Of the 2,376 institutions that used any automated system, 52.6% sent more than 90% of their automated decisions through a single one, and 35.9% sent more than 99%. For roughly a third of reporting lenders, one provider evaluates essentially every automated credit decision they make.

This is the exposure structure P7 describes. Whether an upstream change to either system produces the synchronized loss of historical reproducibility that P7 predicts is not tested here, and the data do not support that test: HMDA records the system used, not its version, and reports the applicant's debt-to-income ratio as filed rather than as evaluated at decision time.