Verification Reversal: Cascades and Synthetic Productivity in an AI-Mediated Economy

Vanessa Beck
M.S. Analytics - ML Specialization, University of Illinois Chicago
Independent Researcher
vanessa.beckk1@gmail.com

ORCID: 0009-0008-6611-535X
GitHub: stochastic-sisyphus

January 2026

Abstract

Generative AI has inverted a relationship that economic measurement systems assumed was stable: the cost of producing an artifact now falls below the cost of verifying it. When verification becomes the binding constraint, rational agents stop checking and start forwarding.

This paper formalizes **verification reversal** as a regime condition and derives its equilibrium consequences. In a sequential forwarding game, we show that when the cost gap between verification and forwarding exceeds the expected benefit of catching errors, a **cascade equilibrium** emerges: agents propagate artifacts regardless of private beliefs, and information aggregation fails even as throughput metrics soar (*Proposition 1*). We then show that market selection, operating on observable throughput rather than latent verification stocks, systematically favors low-verification strategies during stable periods—the very periods that dominate expected duration (*Proposition 2*).

The resulting divergence between measured and verification-adjusted productivity constitutes **synthetic productivity**: conventional TFP rises while utility-relevant output stagnates, because verification effort and remediation burdens are omitted from the measurement frame. We formalize **epistemic debt** as a stock variable—the accumulated gap between system complexity and cognitive grasp—and show how it compounds when verification capacity erodes faster than artifact volume grows.

A distinct contamination channel compounds these dynamics. As model-generated content enters the substrates used for evaluation and decision-support, measurement systems become **endogenously self-referential**. We derive conditions under which this contamination introduces directional bias and extends recognition lags, allowing genuine degradation to persist undetected. The resulting verification bottleneck also creates an exploitable **attack surface**, enabling adversarial artifact injection at reduced detection cost.

We examine two candidate self-correction mechanisms—recursive AI verification and market selection—and identify structural conditions under which both fail. Recursive verification lacks independent rejection signals when models share training distributions; market selection operates on lagging proxies that favor low-verification strategies until crises force revaluation.

Finally, we propose an empirical agenda with instrumentation baselines, anchored by a GitHub

pull-request testbed, to measure verification intensity, remediation burden, cascade fragility, substrate contamination, and the accumulation of epistemic debt. The framework yields eight testable hypotheses with explicit falsification conditions.

Keywords: Information cascades; Verification costs; Generative AI; Total factor productivity; Endogeneity; Epistemic debt; Adversarial robustness; Market selection.

JEL Classification: D83 (Search; Learning; Information); D85 (Network Formation and Analysis); O33 (Technological Change); E01 (Measurement and Data on National Income); L15 (Information and Product Quality).

Paper status: Conceptual framework with testable predictions; full microfoundations are a planned extension.

---

1 Introduction

There is a particular kind of confidence that comes from watching a machine produce, in seconds, what once took hours. Code that compiles. Prose that scans as expert. Forecasts dressed in last quarter's familiar format. Through the lenses we inherited, this reads as productivity: clean, unambiguous, almost divinely efficient.

But efficiency measured against what? And verified by whom?

Measurement systems encode assumptions about what they measure. One assumption sits quietly inside most productivity statistics: that producing an artifact requires cognitive work roughly proportional to the artifact's complexity. More output meant more thinking. The link was never perfect, but it was stable enough to support decades of economic inference.

Generative AI has severed that link.

This paper is an attempt to build a measurement frame for a phenomenon that is still becoming visible. That is an uncomfortable position. The standard move would be to wait: let the data accumulate, let consensus form, then publish the retrospective analysis that confirms what everyone already suspected. But verification reversal, if it is real, degrades the very mechanisms that would eventually surface it. The recognition lag is part of the phenomenon.

So this is provisional. The formal sections will look like theory; the empirical sections will look like a research agenda. The honest framing is that this is a bet (a structured bet, with falsifiable predictions, but a bet nonetheless). I am trying to describe a regime shift while standing inside it, which means I cannot be certain I am not simply pattern-matching on noise. The alternative-waiting for certainty-is not available if the argument is correct.

The marginal cost of *producing* a plausible artifact (code, analysis, report, forecast) has collapsed. The marginal cost of *verifying* whether that artifact is correct, reality-tracking, and safe to act on has not. In many domains it has risen, because the artifacts now require scrutiny they once did not, and the people capable of providing that scrutiny are being asked to do it at scale.

This paper studies what happens when generation becomes cheaper than verification. The result is not merely "more errors." It is an equilibrium shift. Agents stop verifying. Cascades form. Measured productivity drifts away from verification-adjusted productivity. And the mechanisms that supposedly self-correct (recursive AI verification, market selection) are structurally weakened by the same dynamics they would need to reverse.

1.1 Paper Structure and Claims

This paper proceeds as follows:

Model (Sections 3–4). We formalize verification reversal as a cost inequality and derive a sequential forwarding game. *Proposition 1*: When verification costs exceed forwarding costs sufficiently, cascade equilibria emerge in which rational agents propagate unverified artifacts regardless of private beliefs.

Goodhart Mechanism (Section 3.2). We show why throughput becomes a target and verification becomes latent, collapsing the measurement regime under incentive pressure.

Productivity Measurement (Sections 5–7). We distinguish measured from verification-adjusted productivity and formalize epistemic debt as a stock variable. *Claim 1*: Under verification reversal, measured TFP can rise while utility-relevant productivity stagnates (synthetic productivity). *Claim 2*: Model-generated content can contaminate measurement substrates, extending recognition lags.

Self-Correction Failure (Section 8). We analyze recursive verification and market selection. We argue both mechanisms are weakened structurally because the cascade equilibrium is self-reinforcing on the signals that would guide correction.

Empirical Agenda (Section 9). We propose testable hypotheses with instrumentation baselines.

1.2 Key Definitions

Status of claims (evidential hierarchy)

- **Proposition 1 (Section 4.2):** formally derived from model primitives and payoff inequalities (proof sketch included).
- **Claims 1–2 (Sections 5–6):** derived implications under stated reduced-form assumptions. These are conditional predictions.
- **Selection mechanism (Section 8.3):** a verbal hypothesis with a schematic model and clear modeling gaps (entry/exit and capital isn't fully specified).
- **Predictions / hypotheses (Section 9):** operationalizations intended to make the framework falsifiable.

Definition (Verification Reversal). The regime in which the marginal cost of producing an artifact is lower than the marginal cost of verifying it, formally: $\partial C_p/\partial \mathrm{Y} < \partial C_v/\partial \mathrm{Y}$.

Definition (Synthetic Productivity). The appearance of output growth (rising measured TFP) when verification-adjusted, utility-relevant productivity stagnates or declines.

Definition (Verification-Adjusted Productivity). TFP computed using only verified artifacts, discounting unverified volume.

Definition (Epistemic Debt). The accumulated gap between system complexity and cognitive grasp: the stock of artifacts an organization relies upon but does not fully understand.

Definition (Endogeneity Share). The fraction of variance in key regressors attributable to model-generated content: $\alpha_t = \mathrm{Var}(X_m\mathrm{odel,t}) \,/\, \mathrm{Var}(X_t\mathrm{otal,t})$.

Definition (Attack Surface). The set of verification bottlenecks and cascade entry points exploitable by adversaries seeking to inject malicious or misleading artifacts into propagation chains.

1.3 Contributions

This paper offers six contributions:

1. **A forwarding game under verification reversal** in which agents rationally propagate unverified artifacts, generating information blockage even as throughput metrics soar.
2. **A formal distinction between measured and verification-adjusted productivity**, showing how the gap can grow over time as verification capacity erodes.
3. **An endogeneity share parameter** measuring how model-mediated content enters measurement substrates.
4. **Structural conditions under which self-correction fails**: recursive AI verification lacks independent rejection signals when models share training distributions (*Section 8.1–8.2*), and market selection systematically favors low-verification strategies during stable periods (*Proposition 2, Section 8.3*). We also identify verification reversal as an **attack surface** that adversaries can exploit (*Section 10.4*).
5. **A concrete empirical agenda** with testable hypotheses and instrumentation baselines.
6. **An adversarial framing** showing how verification bottlenecks create exploitable structure, with a corresponding empirical hypothesis (*H8*) for measuring adversarial injection success as a function of verification intensity.

1.4 Relation to Existing Literatures

The analysis braids together several threads: foundational work on informational cascades and social learning (Bikhchandani et al., 1992; Banerjee, 1992), emerging work on AI productivity measurement, econometric treatments of endogeneity (Wooldridge, 2010), and philosophical work on epistemic opacity and epistemic debt. Two additional influences structure the argument: Marvin Minsky's *Society of Mind* (1986), which frames robust intelligence as requiring diverse, independently-failing processes; and Hyman Minsky's *Financial Instability Hypothesis* (1986), which describes how stability itself breeds the conditions for crisis.

What distinguishes this framework. The novelty lies in three interacting mechanisms: (i) verification capacity is endogenous and erodes under sustained underuse, so organizations lose not just the time but the *ability* to verify; (ii) measurement substrates can become self-referential as model-generated content enters the data used to detect problems, extending recognition lags; and (iii) market selection operates on lagging proxies (throughput, velocity) rather than latent verification stocks, systematically selecting against high-verification strategies during stable periods. Erosion accelerates contamination, contamination extends detection lags, and selection pressure accelerates erosion.

1.5 How to Read This Paper

This paper occupies a middle ground between formal theory and measurement design. Sections 3–4 develop a stylized model with explicit assumptions and one formally derived proposition. **Section 8.3 extends the formal analysis to market selection dynamics (Proposition 2).** Sections 5–8 present Claims 1–2 and the self-correction analysis as structured arguments that follow from the model if its premises hold. **Section 10.4 develops the adversarial implications.** Section 9 translates predictions into testable hypotheses with concrete operationalizations.

1.6 Empirical Motivation

Three domains exhibit verification reversal dynamics with measurable consequences.

Software development under AI assistance. GitClear's analysis of 153 million changed lines of code across 2020–2024 documents systematic quality shifts as AI code-generation tools scaled. Their 2024 report finds that short-horizon churn (lines reverted within two weeks of being written) increased substantially as AI-assisted commits grew, while "moved" code (a proxy for thoughtful refactoring) declined relative to copy/paste behavior (GitClear, 2024).

Security vulnerability proliferation. Multiple studies document elevated vulnerability rates in AI-generated code, with common weakness enumerations (CWEs) such as SQL injection, XSS, and hardcoded credentials appearing at higher frequencies than in human-written baselines (Pearce et al., 2022). These results are consistent with a verification-reversal mechanism: generation scales cheaply, while deep security review (threat modeling, misuse-case analysis, integration testing) does not.

Knowledge work and analysis. Qualitative reports from consulting, finance, and legal work describe a shift toward "light edit and forward" behavior, with deep verification increasingly reserved for a small share of high-stakes artifacts. Systematic measurement remains sparse, but the structure is the same: artifact volume rises while per-artifact verification time declines.

These cases share the same incentive geometry: throughput is visible, verification is costly, and the penalty for being wrong is lagged and often externalized.

---

2 Literature Review

2.1 Informational Cascades and Social Learning

The mathematics of herding was formalized before anyone imagined machines that could produce plausible text at scale. Bikhchandani, Hirshleifer, and Welch (1992) and Banerjee (1992) showed that rational agents observing predecessors' actions can enter cascades where private signals are discounted and learning stops, even when private information remains informative. The mechanism is elegant: once the inferred weight of upstream behavior exceeds the informational value of one's own signal, the rational move is to follow the crowd. Truth becomes irrelevant to the equilibrium.

These models treat observation costs and signal acquisition costs as primitives. This paper adapts the framework to a setting where generative AI has altered the cost structure in a specific way: verification has become expensive relative to production. The cascade dynamics remain, but the entry conditions have changed. When producing an artifact costs less than checking it, the threshold for cascade formation drops.

2.2 AI Productivity Measurement

Empirical studies document productivity gains from generative AI across knowledge-work tasks. Brynjolfsson et al. (2023) find that customer service agents using AI assistants resolve 14% more issues per hour, with gains concentrated among less-experienced workers. Peng et al. (2023) report that GitHub Copilot users completed coding tasks 55% faster in a controlled experiment. Noy and Zhang (2023) find that ChatGPT assistance reduced writing time for professional tasks by 40% while improving output quality as rated by evaluators.

The gains are real within the measurement frames employed. But the measurement frame matters.

Most studies emphasize output volume and short-horizon quality metrics without separately tracking verification bandwidth, downstream error remediation, or long-run maintenance costs. A programmer who ships twice as many lines of code per day has doubled productivity, unless those lines require three times as much review, generate twice as many bugs, and create maintenance burdens that compound for years.

A key gap in the literature is the absence of verification-adjusted productivity measures. The framework developed here provides a rationale for such adjustments and predicts that conventional productivity metrics systematically overstate welfare gains when verification costs dominate.

2.3 Endogeneity in Econometrics

Econometric theory has long grappled with endogeneity arising from omitted variables, measurement error, simultaneity, and selection. Textbook treatments (Wooldridge, 2010) provide the technical machinery for detection and correction when the problem is recognized.

In AI-mediated economies, a distinct channel emerges: regressors can themselves be model-generated and optimized for plausibility, breaking orthogonality between regressors and errors in ways that standard diagnostics may not detect.

2.4 Epistemic Opacity and Epistemic Debt

Work on epistemic opacity emphasizes that reliability must be argued through verification, validation, robustness testing, and track record. When systems become too complex for any single agent to comprehend, trust must be distributed across institutional processes. Those processes require maintenance.

In software development and AI-assisted work, *epistemic debt* describes how artifact production can outpace human comprehension, creating a growing divergence between system complexity and cognitive grasp. The concept parallels technical debt, but the currency is understanding rather than code quality.

---

3 When Generation Becomes Cheaper than Verification

The defining condition is simple to state and difficult to escape once it binds.

Let Y denote artifact volume per period. Let $C_p(\mathrm{Y})$ and $C_v(\mathrm{Y})$ be the costs of producing and verifying that volume.

Definition (Verification Reversal). The verification reversal regime begins when:

$\partial C_p/\partial \mathrm{Y} < \partial C_v/\partial \mathrm{Y}$

The inequality need not hold universally. It suffices that for a substantial class of "good enough" artifacts (the kind that clear immediate review, satisfy surface criteria, and raise no obvious flags) incremental production is easier than incremental verification.

3.1 Scope Conditions: Where Verification Reversal Binds

This paper is **not** claiming that verification is always expensive or always human-bound. The regime is defined by *artifact classes* where cheap automated verification does not bind.

Artifact classes where cheap automated verification often does not bind:

- **Open-ended analysis artifacts:** memos, forecasts, policy briefs, research notes, "exec-ready" narratives.
- **Socio-technical decision artifacts:** requirements, incident retros, risk assessments, compliance narratives.
- **Software artifacts beyond compile/test:** architecture changes, security properties, long-run maintainability, integration behavior.
- **Measurement substrates:** dashboards, metrics pipelines, benchmark construction, and documentation that later become ground-truth inputs.

Artifact classes where cheap automated verification can bind (limiting cases):

- Artifacts with **tight formal contracts** (types, proofs, model checking).
- Artifacts with **high-coverage automated tests** and fast ground-truth feedback.
- Low-volume, high-stakes domains with enforced audit gates.

3.2 Verification Cost Non-Linearity

Verification cost exhibits non-linearities. As artifact volume Y scales, verification cost per artifact can rise due to cognitive fatigue and needle-in-a-haystack dynamics. Define:

$C_v(Y) = c_0 Y + c_1 Y^{\alpha} + c_2 f(Y/H)$

where $c_0 Y$ is baseline linear cost, $c_1 Y^{alpha}$ with $\alpha > 1$ captures fatigue effects, and c_2f(Y/H) captures increasing error-detection difficulty as signal-to-noise declines (H is human verification bandwidth).

(The notation is clean. The reality is not. What c_2f(Y/H) actually measures is the feeling of a reviewer at 4pm on Friday, facing a queue that grew faster than the hours in the day.)

3.3 The Goodhart Linkage: Throughput as Target, Verification as Latent

When **throughput metrics** (Y, velocity, tickets closed, artifacts shipped) become targets, they stop being useful measures of welfare-relevant productivity.

Under verification reversal, verification quality (v) is typically **latent** and remediation burden (R) is **lagged**. This induces a Goodhart-style collapse of the measurement regime:

- Managers optimize the observable: Y.
- The system draws down the unobserved: v and verification skill.
- The bill appears later: R (rework, incidents, reversions), sometimes only at crisis.

Interpretation: verification erosion is not just "mistakes happen." It is a measurement regime failing under incentive pressure.

3.4 Measurement Gap: Verification Costs vs. Verification Outcomes

The verification cost function $C_v(\mathrm{Y})$ is central to this framework, yet direct measurement of verification *effort* remains sparse. Available evidence documents verification *outcomes* (error rates, churn, defect density) rather than verification *inputs* (reviewer time, cognitive load, attention allocation).

What we can measure:

- **Outcome proxies:** Code churn and revert rates (GitClear, 2024), defect density post-deployment, incident frequency and severity, time-to-detection for known vulnerabilities.
- **Coarse input proxies:** Review queue depth, time-from-submission-to-approval, number of review cycles per artifact.

What we cannot easily measure:

- **Cognitive effort per review:** How much attention does a reviewer allocate to each artifact? Does this decline as volume increases?
- **Verification depth:** Is the reviewer checking surface features (syntax, formatting, obvious errors) or deep properties (logic correctness, security implications, integration effects)?
- **Fatigue dynamics:** How does verification quality degrade across a review session? Across a week? Across exposure to thousands of similar artifacts?

Implication for this framework. The non-linearity in $C_v(\mathrm{Y})$ (specifically the fatigue term $c_1 Y^{alpha}$ with $\alpha > 1$) is theoretically motivated but empirically underdetermined. The framework predicts that verification cost per artifact rises with volume; falsifying this prediction requires direct measurement of verification effort, not just outcomes.

A feasible measurement approach. Time-tracking at the artifact level (review duration per PR, per document, per forecast) combined with quality tagging (substantive comments vs. rubber-stamp approvals) can approximate verification intensity. Eye-tracking and cognitive load measures exist in laboratory settings but are difficult to deploy at organizational scale.

This measurement gap is a limitation of the current literature, not a weakness specific to this framework. Addressing it is part of the empirical agenda proposed in Section 9.

3.5 The Non-Linearity Assumption

The cost function $C_v(\mathrm{Y}) = c_0 Y + c_1 Y^{alpha} + c_2 \mathrm{f}(\mathrm{Y/H})$ is theoretically central to this framework. Yet the non-linearity (specifically whether fatigue effects with $\alpha > 1$ dominate specialization and tooling gains) remains empirically underdetermined.

This is not a weakness we are hiding. It is an assumption doing explicit load-bearing work.

We proceed assuming $\alpha > 1$ as a working hypothesis, grounded in qualitative accounts of review queue dynamics and documented quality degradation under AI-assisted scaling (GitClear, 2024), but not yet directly validated. The assumption is plausible: cognitive fatigue compounds across review sessions, needle-in-haystack dynamics worsen as artifact volume grows, and tooling gains face diminishing returns when the binding constraint is human attention rather than mechanical throughput.

But plausibility is not proof.

What the assumption requires: Direct measurement of whether $\log(\text{review\_time}) \sim \alpha \cdot \log(\text{volume})$ exhibits $\alpha > 1$ across organizational contexts. If $\alpha \leq 1$ (if specialization and tooling gains

outpace fatigue effects), then the cost structure may not reverse at organizational scale, and cascade formation becomes contingent on other factors (incentive misalignment, principal-agent dynamics) rather than rational response to cost structure.

Falsification condition (F1): Regress log(review_time_per_artifact) on log(artifact_volume) within organizations. If slope ≤ 1, the non-linearity does not bind as described.

Until this test is run, the framework rests on a contestable premise about cost structure. This is its primary empirical vulnerability, and we are naming it explicitly rather than obscuring it in notation.

---

4 A Forwarding Game and Rational Cascades

Consider a workflow as a sequence of agents: i = 1, 2, . . . , N. An artifact enters at one end and passes through hands until it exits as a decision, a deployment, a publication, a commitment. At each position, the agent faces a choice:

- **Verify:** incur cost C_v to check the artifact against independent criteria.
- **Forward:** pass it on with minimal scrutiny at cost C_f, where C_f « C_v.

4.1 Model Primitives

State space. Each artifact has a true state $\omega \in$ {valid, invalid}, drawn with prior P(valid) = q_0.

Information structure. Agent i receives a private signal s_i $\in$ {good, bad} with precision P(s_i = good | valid) = P(s_i = bad | invalid) = p > 0.5. Agent i also observes the sequence of actions a_1, . . . , a_{i-1}.

Actions. Agent i chooses a_i $\in$ {Verify, Forward}.

Payoffs. If agent i verifies:

- Cost: C_v
- Benefit: if the artifact is invalid and caught, agent i receives B (avoiding downstream error costs)

If agent i forwards:

- Cost: C_f
- If the artifact is invalid and causes downstream harm, agent i bears expected cost λD where λ is the probability of attribution and D is reputational/organizational damage.

4.2 The Forwarding Condition

Agent i forwards if verification's expected net benefit is below its cost gap:

C_v - C_f > (1 - μ_i) · (B - λD)

$LetB_net = B - \lambda D.Rearranging$:

Proposition 1 (Forwarding Threshold). Agent i forwards if and only if:

$\mu_i > 1$ - $(C_v$ - $C_f)$ / $B_n$et

When C_v - $C_f \geq B_n$et, the threshold is non-positive, so the condition holds for all $\mu_i \in [0,1]$: agents forward regardless of beliefs.

Proof sketch. Follows directly from expected utility comparison.

Cascade formation. Once forwarding becomes optimal regardless of s_i, the action conveys no information. The cascade is absorbing: once entered, it cannot be exited by any single agent's deviation.

4.3 Information Blockage

Definition (Information Blockage). A state in which observed actions are uninformative about artifact validity: the mutual information between the action sequence (a_1, ..., a_n) and the true state ω approaches zero, even as throughput remains high.

The cascade is absorbing. This is the technical term for what it feels like to watch a document you know is wrong get approved by three layers of review because no one had time to read it.

This information blockage mechanism connects to foundational work on adverse selection under quality uncertainty (Akerlof, 1970): when verification costs make quality unobservable, low-quality artifacts can drive out high-quality ones through a forwarding dynamic rather than a pricing dynamic.

---

5 Synthetic Productivity and Regime Misidentification

At the macro level, total factor productivity (TFP) is typically computed as:

$TFP^{m}_{t}easured = Y_t/F(K_t, L_t)$

5.1 A Canonical Productivity Wedge Estimand

Let:

- Y be gross artifact output (throughput).
- v $\in [0,1]$ be the effective verification rate.
- R be remediation cost (rework, incident response, rollback, downstream correction).

Definition (Net Verified Output). Net Verified Output := Y · v - R

Definition (Synthetic Productivity). Δ := Y - (Y · v - R)

Under verification reversal, Y can grow independently of verified value creation.

Claim 1 (Synthetic Productivity). Under volume–value decoupling, measured TFP rises while verification-adjusted TFP stagnates or declines.

---

6 Endogenous Measurement: Models Eating Their Own Outputs

Econometric analysis often assumes:

$Y_t = \beta X_t + \epsilon_t with E[X_t \epsilon_t] = 0$

In an AI-mediated economy, the data used to construct X_t increasingly include model-generated artifacts. Write:

$X_t = X_h, t + X_m, t$

where X_h,t is human- or sensor-grounded content and X_m,t is model-mediated content.

OLS bias is driven by the correlation term, not by variance share alone:

$\beta_O LS = \beta + Cov(X_t, \epsilon_t)/Var(X_t)$

6.1 Endogeneity Share (Exposure)

Definition (Endogeneity Share). α_t = Var(X_m,t) / Var(X_t)

α_t measures **exposure**: how much of the substrate is model-mediated. It is not itself bias.

6.2 When Exposure Becomes Bias (Directionality Conditions)

Bias emerges when model-mediated content is systematically wrong in ways correlated with outcomes:

Cov(X_m,t, ϵ_t) $\neq 0$

This occurs under at least three concrete mechanisms:

1. **Optimization for acceptance under proxy incentives.** When organizations reward "looks right" or "clears review," model outputs are selected for plausibility, not truth. Selection pressure can induce directional errors aligned with the proxy objective (e.g., smoothing toward the expected trend; reducing variance; overconfidence in favored narratives).
2. **Benchmark and evaluation gaming / contamination.** If X includes evaluation artifacts (benchmarks, rubric-scored outputs, "quality" labels) that models indirectly influence or learn, then X_m,t inherits the model's inductive biases. Errors become correlated with ϵ_t because the same model family shapes both regressor construction and the underlying process generating outcomes.
3. **Distributional blind spots.** Under shifts where tail events matter (rare failures, security vulnerabilities, adversarial behavior), models can be directionally wrong in the tails. If outcomes ϵ_t are driven by tail events, then X_m,t (which smooths tails) becomes correlated with ϵ_t.

6.3 Worked Example: Earnings Forecast Contamination

Consider a stylized earnings forecasting pipeline where analysts produce quarterly EPS estimates.

Setup. Let true earnings be:

$E_t = \mu + \epsilon_t$ where $\epsilon_t \sim N(0, \sigma^2)$

Analysts observe signals and produce forecasts F_t. Historically, F_t was human-generated with:

$F_t^{human} = E_t + \eta_t$ where $\eta_t \sim$ N(0, τ^2) and Cov(η_t, ϵ_t) = 0

$The regressor X_t = F_t was unbiased : E[F_t] = E[E_t] = \mu.$

Model contamination. Now suppose a fraction α of forecasts are model-generated:

$F_t^{model} = \beta \cdot X_{\{t-1\}} + \nu_t$

where X_{t-1} includes prior model forecasts (autoregressive contamination) and β is estimated from recent data that increasingly contains model outputs.

The composite forecast becomes:

$F_t = (1 - \alpha) \cdot F_t^{human} + \alpha \cdot F_t^{model}$

Bias emergence. Under proxy optimization (Mechanism 1), models learn that forecasts closer to consensus clear review faster. This induces:

E[$F_t^{model} \mid E_t < \mu$] > E_t (upward bias when true earnings are low)

E[$F_t^{model} \mid E_t > \mu$] < E_t (downward bias when true earnings are high)

The model compresses toward μ, reducing variance but introducing systematic directional error.

Numerical illustration. Let $\sigma^2 = 1$ (true earnings variance), $\tau^2 = 0.25$ (human forecast noise), and suppose model forecasts exhibit compression bias (a directional error pattern consistent with optimization for acceptance rather than accuracy):

$F_t^{model} = 0.7 \cdot E_t + 0.3 \cdot \mu + \nu_t$

Then:

$Cov(F_t^{model}, \epsilon_t) = 0.7 \cdot Var(\epsilon_t) = 0.7 \neq 0$

The model-generated component is correlated with the error term because it systematically underweights tail realizations.

Bias in regression. If we regress realized earnings on forecasts:

$E_t = \beta \cdot F_t + u_t$

OLS bias is:

$\beta_{OLS} - \beta = Cov(F_t, u_t)/Var(F_t)$

With $\alpha = 0.3$ (30% model-generated forecasts) and the compression structure above:

$Cov(F_t, u_t) = \alpha \cdot Cov(F_t^{model}, \epsilon_t) \approx 0.3 \cdot 0.7 = 0.21$

This is non-negligible bias arising purely from substrate contamination, even though human forecasts remain unbiased.

Detection difficulty. Standard specification tests (Hausman, Durbin-Wu) require valid instruments. But if alternative data sources (analyst reports, news summaries, management guidance) are themselves increasingly model-mediated, instrument validity degrades. The bias becomes difficult to detect because the diagnostic tools rely on substrates that share the contamination.

6.4 Recognition Lag Extension

Let τ be expected time-to-detection of genuine misalignment between model-mediated measurement and reality.

Claim 2 (Recognition Lag Extension). Holding true degradation fixed, increasing α_t can increase τ when model-mediated substrates suppress residual visibility or reduce the probability of collecting independent ground truth.

Interpretation: as the measurement apparatus becomes self-referential, problems take longer to surface and have more time to compound.

6.5 Two Lag Mechanisms: Detection vs. Occurrence

Claim 2 asserts that recognition lags extend as model-mediated substrate share (α_t) rises. This claim conflates two distinct mechanisms that require separation.

Definition (Detection Lag). Let τ_d be the expected time from problem occurrence to problem detection, holding occurrence rate constant.

Definition (Occurrence Lag). Let τ_o be the expected time from initial condition to problem occurrence, holding detection capacity constant.

These are not the same. Detection lag is a *measurement* problem: the underlying degradation rate is unchanged, but we find problems slower because diagnostic tools are contaminated. Occurrence lag is a *dynamics* problem: problems emerge more slowly initially (because contaminated models smooth tails and suppress variance), but compound faster and erupt at larger scale when they finally manifest.

Drivers of each mechanism:

Detection lag (τ_d) increases when:

- α_t rises (more substrate is model-mediated)
- Diagnostic tools rely on the contaminated substrate
- Independent ground-truth collection declines

Occurrence lag (τ_o) increases when:

- Model outputs compress toward consensus (tail suppression)
- Variance in measured outcomes declines
- Early warning signals are smoothed away

Different empirical predictions:

Mechanism	Prediction	Observable
Detection lag	Fixed underlying error rate, slower discovery	Time-to-detection for seeded de
Occurrence lag	Lower observed error rate initially, larger eventual failures	Variance in outcomes declines, t

Clarification of Claim 2:

Claim 2 as stated (that increasing α_t can increase τ, or recognition lag) is a claim about *detection lag*. It asserts that model-mediated substrates suppress residual visibility and reduce the probability of collecting independent ground truth.

The occurrence lag mechanism is a separate, complementary claim: that tail suppression in model outputs delays problem manifestation while allowing underlying fragility to compound. Testing this requires tracking not just time-to-detection but the *severity distribution* of detected problems over time.

Joint prediction: If both mechanisms operate, we should observe:

1. Declining variance in routine metrics (occurrence lag)
2. Increasing time-to-detection for known anomalies (detection lag)
3. Increasing severity of detected problems when they finally surface (compounding)

This is testable. Track variance, detection latency, and severity jointly. If variance declines and detection latency increases but severity remains constant, only detection lag operates. If severity increases as variance declines, both mechanisms operate.

---

7 Verification-Adjusted Productivity and the Verification Treadmill

Define verification-adjusted utility as:

$V(Y, \theta) = \int_0^Y v(y, \theta) dy$

Measured vs. verification-adjusted productivity:

$TFP_t^measured = Y_t / F(K_t, L_t) TFP_t^actual = V(Y_t, \theta_t) / F(K_t, L_t)$

7.1 The Verification Treadmill

A reduced-form representation:

$v(Y, \theta) = \theta \cdot g(Y/\theta), g'(\cdot) < 0$

Prediction 1 (Wedge Growth). If $\mathrm{d}Y_t/\mathrm{dt} > 0$ while $\mathrm{d}\theta_t/\mathrm{dt} \leq 0$, then $\mathrm{d}/\mathrm{dt}(\mathrm{TF}P_t^{measured}$ - $\mathrm{TF}P_t^{actual}) > 0$.

7.2 Epistemic Debt as a Stock

Let C_s(t) be complexity (or volume) of epistemically loaded artifacts and G_c(t) be cognitive grasp (a function of verification capacity θ_t and verification skill κ_t).

Definition (Epistemic Debt). $D_e(\mathrm{T}) = \int_0^T$ (C_s(t) - G_c(t)) dt

Note: Epistemic debt as defined here is a conceptual stock variable, not a directly measured quantity. C_s(t) can be approximated by artifact volume and integration complexity; G_c(t) cannot be directly observed but can be proxied by learning outputs (post-mortem closure rates, incident recurrence, process adaptation). The integral formulation is a modeling device, not a measurement claim.

The metaphor is useful precisely because debt implies a creditor. The creditor, in this case, is reality. It does not negotiate.

7.3 Synthetic Productivity as a Drawdown of Epistemic Capital

A separate, compounding channel matters for the long-run measurement substrate.

If successive model generations are trained on model-generated (synthetic) data, the system can *consume* the epistemic capital embedded in the historical corpus of human-generated, reality-anchored information. Shumailov et al. (2024) show that indiscriminate recursive training on model-generated content induces "model collapse," with low-probability tails disappearing and learned behavior converging toward degenerate distributions over generations.

Within this framework, the mapping is direct:

- Higher α_t (model-mediated substrate share) increases the probability that "ground truth" inputs become self-referential.
- Self-referential substrates reduce tail sensitivity and anomaly visibility.
- Reduced tail sensitivity increases the likelihood that verification reversal persists, because fewer failures are detected early enough to trigger reinvestment in verification capacity.

Interpretation: synthetic productivity can look like a permanent efficiency gain while it is quietly drawing down a finite stock of epistemic capital (and making future verification harder).

---

8 Why Self-Correction Mechanisms Fail

Two self-correction arguments dominate optimistic discourse. The first: AI will verify AI, lowering verification costs. The second: market selection will weed out low-verification strategies.

Both mechanisms are structurally weakened when verification reversal holds, because the cascade equilibrium is self-reinforcing on the signals that would guide correction.

8.1 Recursive Verification and Correlated Error

Effective verification requires an independent rejection signal.

Definition (Independent Rejection Signal). A verification procedure V provides an independent rejection signal for generator G if, conditional on G's surface features, V's rejection event is not driven by those same features.

Cross-model error correlation (ρ) is a proxy for independence failure.

8.2 The Double Minsky Framework

Marvin Minsky emphasizes robustness through diverse, independently failing processes. Hyman Minsky emphasizes that stability breeds instability. Under verification reversal, recursive AI verification can create fluent consensus without truth-sensitivity, while stable-looking throughput regimes draw down verification capacity and build epistemic leverage.

I did not set out to name a framework after two people named Minsky. The parallel emerged in conversation, as parallels do. It stuck because it was true.

8.3 Market Selection on Lagging Indicators

The intuition is straightforward: if markets allocate resources based on observable throughput (π) rather than latent verification stock (θ), then low-verification strategies will be systematically favored during stable periods. This section formalizes that intuition.

Model Setup

Consider a population of organizations indexed by type $\tau \in \{\text{H, L}\}$:

- **Type H (high verification):** Chooses (Y_H, θ_H) with Y_H < Y_L and $\theta_H > \theta_L$
- **Type L (low verification):** Chooses (Y_L, θ_L) with Y_L > Y_H and $\theta_L < \theta_H$

Let m_t ∈ [0,1] denote the market share of Type H at time t.

Payoff Structure

In stable periods (no crisis), observable performance determines resource allocation:

$\pi_\tau = f(Y_\tau) where f' > 0$

Type L dominates on observable performance during stable periods: $\pi_L > \pi_H$.

Crises arrive stochastically with hazard rate:

$h_t = h_0 \cdot g(D_{\{e,t\}}/\theta_t)$

where g' > 0, so crisis probability increases with epistemic leverage (debt-to-verification ratio).

In crisis, Type τ suffers loss:

$L_\tau = L_0/\theta_\tau$

with L_H < L_L because higher verification stock buffers against crisis losses.

Market Share Dynamics

During stable periods:

$dm_t/dt|_{\{stable\}} = -s \cdot (\pi_L - \pi_H) \cdot m_t \cdot (1 - m_t)$

where s > 0 is selection intensity. Type H loses share because it is outperformed on observable metrics.

During crisis:

$dm_t/dt|_{\{crisis\}} = +c \cdot (L_L - L_H) \cdot m_t \cdot (1 - m_t)$

where c > 0 is crisis-induced reallocation intensity. Type H gains share because it suffers lower losses.

Proposition 2 (Selection Against Verification)

Let T_s be expected duration of stable periods and T_c be expected duration of crises. Over a complete cycle of length $T_s + T_c$:

$E[\Delta m] = m(1 - m) \cdot [- s(\pi_L - \pi_H) \cdot T_s + c(L_L - L_H) \cdot T_c]$

If:

s · (π_L - π_H) ·T_s > c · (L_L - L_H) ·T_c

then E[Δm] < 0: market share of high-verification organizations declines in expectation.

Proof. The expected change in market share is the time-weighted sum of selection effects. During stable periods, Type H loses share at rate s(π_L - π_H). During crises, Type H gains share at rate c(L_L - L_H). The net effect depends on the duration-weighted comparison. When stable periods dominate expected duration (T_s » T_c), the stable-period selection effect dominates. ■

Corollary (Crisis Scarcity Amplifies Selection)

As $T_s \to \infty$ (crises become rare), selection against verification intensifies regardless of crisis severity.

Critical Assumptions

This result requires:

1. θ **is latent (unpriced).** Markets cannot observe or price verification capacity pre-crisis. If θ were observable, investors could allocate to high-θ organizations and the selection dynamic would weaken or reverse. This connects to Grossman and Stiglitz (1980): if acquiring information about θ is costly and the information cannot be captured privately, markets systematically underprice it, and the selection mechanism operates on observables (throughput) rather than fundamentals (verification stock).
2. **No entry/exit.** Both types persist; no organization is selected out entirely. If Type H organizations exit during stable periods, the population loses high-verification capacity permanently.
3. **Smooth dynamics.** No threshold effects or catastrophic transitions. If crises cause discontinuous elimination of Type L organizations, the dynamics change.
4. **Exogenous crisis hazard.** The hazard rate h_t depends on aggregate epistemic leverage, not on individual firm behavior. If organizations could individually reduce crisis probability, private incentives might sustain verification.

Boundary Cases Where Selection Weakens

- **Observable θ:** If verification capacity can be audited, certified, or credibly disclosed, markets can price it pre-crisis. Then selection operates on θ directly, not just π.
- **Frequent crises:** If T_c is large relative to T_s, crisis-period selection dominates and high-verification strategies are favored.
- **Catastrophic crisis losses:** If L_L is large enough to eliminate Type L organizations entirely during crises, high-verification types survive by default.
- **Organizational learning:** If Type L organizations can rapidly increase θ when crisis becomes visible, they can mimic Type H during crises and avoid selection.

Interpretation. The selection mechanism is not a market failure in the traditional sense—agents optimize correctly given observable information. The failure is structural: the payoff-relevant variable (θ) is latent, and the selection-relevant variable (π) favors low-verification strategies during the intervals that dominate expected duration.

---

9 Empirical Hypotheses and Instrumentation

Theory without measurement is philosophy. This section operationalizes the framework into testable hypotheses and instrumentation baselines. The goal is not to "prove" verification reversal, but to define measurable objects that could falsify it.

9.1 One MVP Testbed (Immediate Feasibility)

Public GitHub repositories + pull request metadata

- **Y (throughput):** commits, lines changed, PRs merged.
- **v (verification intensity proxy):** review depth (review duration, substantive comments, requested changes), number of revision cycles.
- **R (remediation):** churn/reverts, bug-fix PR share, rollback commits, incident-linked fixes.

This is deliberately minimal: it provides a concrete place to start measuring the wedge implied by synthetic productivity, and it directly connects to existing evidence on AI-assisted codebase quality shifts (GitClear, 2024).

A direct proxy for epistemic debt accumulation: compare Time-to-Resolve (TTR) for defects in AI-assisted versus human-written code, controlling for defect severity. If epistemic debt compounds as predicted, TTR should increase disproportionately for AI-assisted code because maintainers lack the deep understanding required to diagnose and fix logic they did not produce.

9.2 Direct Measurement of Verification Cost Non-Linearity

The non-linearity assumption ($\alpha > 1$ in $C_v(\mathrm{Y}) = \mathrm{c\_0Y} + \mathrm{c\_1}Y^{\alpha}$) is central to the framework. This section proposes a measurement protocol to test it.

Hypothesis: Verification cost per artifact is super-linear in artifact volume: $\alpha > 1$.

Direct Test: Review Time Scaling

Method: Within a set of repositories with comparable complexity, regress log(mean_review_time) on log(P$R_v$olume) across time periods.

Model:$log(T_review) = \beta_0 + \alpha \cdot log(Y) + \epsilon$

Prediction: $\alpha > 1$ (super-linear scaling)

Falsification: $\alpha \leq 1$ (linear or sub-linear scaling)

Data source: GitHub API (PR timestamps, review comments, approval times); internal DevOps logs.

Controls: PR complexity (lines changed, files touched); reviewer experience; repository characteristics.

Indirect Test: Seeded Defect Detection Probability

Method: Inject known defects at controlled rates into codebases with varying PR volume. Measure detection probability as a function of volume.

Model: P(detect) = f(Y, defect_type)

Prediction: P(detect) declines with Y (needle-in-haystack effect)

Falsification: P(detect) is constant or increasing in Y

Data source: Controlled experiment with synthetic bugs.

Indirect Test: Reviewer Fatigue

Method: Track review quality (substantive comments, requested changes, revision cycles) as a function of position in reviewer's daily queue.

Model: Quality_i = g(position_i, total_volume)

Prediction: Quality declines with queue position and total volume

Falsification: Quality is independent of position and volume

Data source: Reviewer activity logs with timestamps.

Limitations:

1. Review time is a proxy for cognitive cost; actual effort is not observed.
2. Seeded defects are artificial; real defect detection may differ.
3. Fatigue effects may be confounded with time-of-day effects.
4. Sample is limited to software development; generalization to other domains requires separate testing.

Interpretation:

If all three tests support $\alpha > 1$, the non-linearity assumption is empirically grounded for software domains. If tests are mixed or fail, the cost structure may not reverse at organizational scale, and the cascade mechanism requires alternative drivers (e.g., incentive misalignment rather than cost asymmetry).

9.3 Testable Predictions (Summary)

1. The wedge between measured and verification-adjusted productivity grows over time (H1).
2. Cascade fragility increases with coupling and chain length (H2).
3. Endogeneity share in measurement substrates rises over time (H3).
4. Feedback loops exhibit learning collapse as incidents stop changing processes (H4).
5. Verification capacity and capability erode under sustained underuse (H5).
6. Organizational language converges toward model-preferred distributions (H6).
7. Irreversible lock-in accumulates through tool deprecation and skill-pipeline changes (H7).

9.4 H1: Divergence Between Measured and Verification-Adjusted Productivity

Hypothesis:

$\mathrm{d}/\mathrm{dt}(\mathrm{TF}P_t^{measured} - \mathrm{TF}P_t^{actual}) > 0$

Operational objects:

- **Throughput (Y):** artifact volume per period.
- **Verification intensity proxy (v):** review depth, audit coverage, validation lag.
- **Correction burden (R):** rework hours, incident remediation, rollbacks, defect density.

Falsification condition: Correction burden grows slower than (or equal to) throughput gains *and* verification intensity remains stable.

9.5 H2: Cascade Fragility Increases with Coupling and Length

Hypothesis:

$\partial^2 \mathrm{Fragility} \,/\, (\partial \mathrm{NetworkDensity} \cdot \partial \mathrm{CascadeLength}) > 0$

Operational objects:

- **Handoff network:** creator → reviewer → approver → deployer.
- **Propagation distance:** hops survived before detection.
- **Coupling proxies:** dependency graph density, shared reviewers, shared templates.

9.6 H3: Rising Endogeneity Share in Measurement Substrates

Hypothesis:

$\mathrm{d}\alpha_t/\mathrm{dt} > 0$

Operational objects:

- **Provenance tagging:** human vs model vs sensor.
- **Endogeneity share:** $\alpha_t = \mathrm{Var}(X_m\mathrm{odel,t}) \,/\, \mathrm{Var}(X_t\mathrm{otal,t})$.

Note: α_t measures exposure. Bias is governed by $\mathrm{Cov}(X_m\mathrm{odel,t}, \epsilon_t)$ (Section 6).

9.7 H4: Learning Collapse in Feedback Loops

Hypothesis: Learning efficiency decays toward zero:

$\mathrm{d}L_t/\mathrm{dt} < 0$ with $L_t \to 0$

where L_t can be proxied by the ratio of durable process/tooling changes to incidents.

9.8 H5: Erosion of Verification Capacity and Capability

Hypothesis:

$\mathrm{d}\theta_t/\mathrm{dt} \leq 0$ and $\mathrm{d}\kappa_t/\mathrm{dt} < 0$

where θ is verification capacity (labor/time/tooling) and κ is verification capability (skill/accuracy).

A concrete, falsifiable capability test is **seeded defect detection** (inject known bugs and track detection rates over time). This directly targets failure modes documented in evaluations of AI-assisted coding security (Pearce et al., 2022).

9.9 H6: Language and Abstraction Convergence

Hypothesis:

$\mathrm{d}\gamma_t/\mathrm{dt} > 0$ and $\partial\mathrm{Fragility}/\partial\gamma > 0$

where γ_t measures coupling between internal language and model-preferred distributions (e.g., template similarity, phrase entropy reduction, boilerplate share).

9.10 H7: Irreversibility and Lock-In

Hypothesis:

$\mathrm{d}L_t/\mathrm{dt} > 0$

where L_t is a lock-in stock driven by tool deprecation, gate removal, and skill-pipeline changes.

9.11 H8: Adversarial Exploitation of Verification Bottlenecks

Hypothesis: Malicious or misleading artifact injection success rate increases as verification intensity decreases:

$\partial(\text{injection\_success}) / \partial\theta < 0$

Operational objects:

- **Red team penetration rate:** Fraction of seeded adversarial artifacts that survive verification and enter production/decision pipelines.
- **Detection latency:** Time from injection to detection for adversarial artifacts, stratified by verification intensity at injection time.
- **Cascade distance for adversarial vs. benign artifacts:** Do adversarial artifacts propagate further than benign artifacts of similar surface quality?

Instrumentation approach:

1. Establish baseline detection rates under current verification regimes using controlled red-team exercises.
2. Correlate detection rates with verification intensity proxies (review time, reviewer load, queue depth).
3. Track detection latency over time as verification intensity changes.

Falsification condition: Detection rates remain stable or improve as verification intensity declines, indicating that verification quality is not capacity-constrained.

9.12 Appendix A: Instrumentation Details

9.13 A.1 H1: TFP Divergence — Detailed Measurement

Artifact throughput measures (Y):

- Reports produced per analyst per week
- Code commits and lines changed per developer
- Ticket closure rate and time-to-close
- Forecast volume and update frequency
- Dashboard creation and modification rates

Validation / verification measures (v proxies):

- Time from artifact creation to first substantive review
- Fraction of artifacts receiving deep review vs. light edit
- Review queue depth and wait times
- Audit completion rates and coverage

Correction burden measures (R):

- Rework hours per initial artifact hour
- Incident remediation costs (labor + downtime)
- Rollback frequency and scope
- Post-deployment defect density

- Customer-reported error rates and severity distribution

Note that time spent reading code vs. writing it typically follows a 10:1 ratio (Martin, 2008); AI assistants optimize the writing phase while potentially increasing cognitive load in verification.

Implementation sketch: Pair production logs with review/correction logs and compute an estimand of the wedge (Δ) over time.

9.14 A.2 H2: Cascade Fragility — Network Analysis

Handoff network construction:

- Map workflow: who creates → who reviews → who approves → who deploys
- Edge weight: frequency and volume of artifact transfers
- Node properties: verification capacity, throughput, position in chain

Propagation distance:

- Inject synthetic errors at known positions
- Track hops before detection
- Compare by error type and artifact class

Correlated failure analysis:

- Trace provenance backward from detected failures
- Test whether adjacent nodes exhibit correlated error patterns

9.15 A.3 H3: Endogeneity Share — Data Provenance Tracking

Lineage flags:

- Tag all data inputs as: human-generated, model-generated, sensor/ground-truth
- Track provenance through transformation pipelines
- Measure the fraction of key regressors with model-generated ancestry

Substrate share calculation:

- $Decompose : X_t = X_human, t + X_model, t + X_sensor, t$
- $Compute : \alpha_t = Var(X_model, t)/Var(X_t)$

Benchmark contamination checks:

- Test whether evaluation datasets contain model-generated content
- Monitor overlap between model outputs and measurement inputs

9.16 A.4 H4: Learning Collapse — Feedback Loop Closure

Postmortem linkage:

- Count incidents → track postmortems → measure policy/process changes
- Compute: Process changes / Incidents over time

Recurrence rate:

- Classify incidents by root cause
- Measure: Repeat incidents / Total incidents

Gate coverage:

- Enumerate validation checkpoints in workflows
- Measure: Artifacts passing through gates / Total artifacts

9.17 A.5 H5: Verification Capacity and Capability — Stock Measurement

Capacity metrics (θ proxies):

- Verification labor hours as share of total production hours
- Headcount in review, audit, QA, and validation roles
- Budget allocated to verification infrastructure
- Tool coverage: fraction of artifacts passing through automated checks

Capability metrics (κ):

- Time-to-competent-review for novel artifact types
- Seeded defect detection rate (inject known errors; measure catch rate)
- Review accuracy under controlled conditions

This is where AI-generated-code security findings become directly testable under an institutional lens (review pipelines as measurement devices), rather than only as model eval artifacts (Pearce et al., 2022).

9.18 A.6 H6: Language Convergence — Semantic Drift Analysis

Jargon overlap:

- Compare internal documents to a reference corpus (proxy for model-preferred phrasing)
- Measure phrase frequency drift and template similarity

Template convergence:

- Track boilerplate share and structural entropy reduction

Acceptance friction:

- Compare review time and rejection rates for model-generated vs human-generated artifacts

9.19 A.7 H7: Irreversibility and Lock-In — Institutional Commitment Tracking

Tool deprecation:

- Track removal of manual verification tools
- Measure availability of non-AI validation pathways

Skill pipeline:

- Track hiring criteria (verification skills vs throughput skills)
- Monitor training investment and promotion patterns in verification functions

Architectural commitments:

- Count removal of validation gates and approval steps
- Track adoption of auto-merge / auto-deploy workflows
- Track integration depth (how many systems depend on model-mediated inputs)

---

10 Discussion: What the Framework Cannot See

Every measurement frame has blind spots. This one is no exception.

The framework predicts that verification reversal creates cascade equilibria, synthetic productivity, and compounding epistemic debt. If correct, it explains a pattern of institutional fragility that would otherwise appear as a series of unrelated failures. If incorrect, it is an elaborate overfitting—a story imposed on noise, itself an artifact of the pattern-completion dynamics it claims to describe.

I cannot rule out the second possibility from inside the frame. This is not false modesty; it is a structural limitation. The same mechanisms that would make verification reversal real would also make it difficult to detect, and would make confident detection claims suspicious. A framework that predicts its own undetectability is not thereby confirmed.

What I can do is specify conditions under which the framework fails, and commit to those conditions as falsification criteria. Section 9 attempts this. The hypotheses are genuine bets: if verification intensity remains stable or improves as artifact volume grows, the core cost asymmetry does not bind. If cascade fragility does not increase with coupling, the propagation mechanism is not as described. If seeded defects are caught at stable rates regardless of volume, verification capacity is not the binding constraint.

The uncomfortable possibility (the one I keep returning to) is that the framework might be approximately correct and still useless. If verification reversal is a stable attractor rather than a transitional disequilibrium, then knowing about it changes nothing. You cannot outspend an architectural asymmetry. The policy response becomes "accept degraded epistemic conditions and build systems that don't depend on verification being abundant," which is less a solution than a managed decline.

I do not know which world we are in. The paper is an attempt to find out.

10.1 Implications for Measurement

Productivity statistics that ignore verification costs and epistemic debt do not measure what they claim to measure. A verification-adjusted framework would track verification capacity as an asset, epistemic debt as a liability, and correction burden as a cost.

10.2 Implications for Institutional Design

Verification capacity should be treated as critical infrastructure: maintained, invested in, protected from quarterly optimization pressures.

10.3 Implications for AI Development

Heterogeneity in models, training distributions, and verification methods reduces correlated failure risk. A monoculture of frontier models trained on similar data and objectives is a fragility multiplier.

10.4 Adversarial Dynamics: Verification Reversal as Attack Surface

The cascade equilibrium and synthetic productivity mechanisms described in Sections 4–7 do not require adversarial actors. They emerge from decentralized optimization against cost structures. This section describes a *distinct* mechanism: how verification reversal creates exploitable structure for sophisticated adversaries.

The connection is enabling, not causal. Verification reversal lowers the cost of successful artifact injection by reducing detection probability. An adversary who would have been caught under high-verification regimes can now succeed simply because the verification bandwidth is exhausted on routine throughput. The attack surface exists because the defense is structurally weakened.

Verification reversal does not merely create inefficiency. It creates exploitable structure. When verification is the binding constraint, adversaries can optimize against it.

The attack geometry. An adversary seeking to inject malicious or misleading artifacts faces two challenges: (i) producing artifacts that appear legitimate, and (ii) surviving the verification process. Under verification reversal, the second constraint relaxes dramatically.

Let p_detect be the probability that a malicious artifact is detected during verification. Under high-verification regimes:

p_detect = f(verification_intensity, artifact_sophistication)

Under verification reversal, verification_intensity declines while artifact_sophistication (for adversaries with access to the same generative tools) can remain high or increase. The detection probability falls:

$dp\_detect/dt < 0$ when $d\theta/dt < 0$ and $d(artifact\_volume)/dt > 0$

Adversarial strategies enabled by verification reversal:

1. **Volume flooding.** Generate high volumes of marginally-plausible artifacts to exhaust reviewer bandwidth, ensuring that genuinely malicious artifacts face depleted verification capacity.

2. **Cascade exploitation.** Target early nodes in forwarding chains. Once an artifact enters a cascade, it propagates regardless of downstream private signals (Proposition 1). Successful early insertion yields high propagation distance.
3. **Legitimacy mimicry.** Optimize artifacts for the surface features that trigger acceptance (formatting, tone, citation patterns, expected structure) while embedding payload in dimensions reviewers are less likely to scrutinize under time pressure.
4. **Measurement substrate poisoning.** Inject content into the data sources used for evaluation, benchmarking, or decision-support. As α_t rises, the adversary's content becomes part of the "ground truth" against which future artifacts are assessed.

Structural asymmetry. Defenders must verify everything; attackers need only one success. Verification reversal amplifies this asymmetry by reducing the defenders' per-artifact verification budget while leaving the attackers' generation capacity unconstrained.

Security implications. Maintaining diverse verification channels (humans with different expertise, models with distinct training distributions, formal methods where applicable, red-team processes with adversarial mandates) is not optional resilience. It is a security control.

The adversarial framing also suggests that verification reversal may be actively accelerated by sophisticated actors who benefit from reduced verification intensity. This is not a claim about current AI development, but a structural observation: any actor who profits from unverified propagation has an incentive to increase artifact volume relative to verification capacity.

10.5 Limitations and Scope Conditions

This framework applies when verification costs exceed production costs for a substantial class of artifacts. It is limited in domains with tight formal contracts, high-coverage automated tests, low-volume high-stakes decisions with enforced audit gates, or fast and unambiguous feedback loops.

The computational substrate also imposes physical constraints. Data center electricity consumption is projected to double by 2026 (IEA, 2024), suggesting synthetic productivity may face resource limits independent of verification dynamics.

The framework's predictions bind most strongly where failure costs are high or externalized, feedback loops are slow or noisy, and remediation requires human comprehension. In domains where failure is cheap, feedback is fast, and remediation is automated—consumer-facing feature iteration with instant rollback, for instance—verification reversal may represent efficient resource allocation rather than institutional pathology. The boundary between these regimes is itself an empirical question.

---

11 Falsification Conditions and Boundary Cases

This section consolidates the conditions under which the framework would be falsified, the tests that would establish each condition, and the boundary cases where predictions weaken.

F1. Cost Asymmetry (Verification Reversal Regime)

Core claim: $\partial C_p/\partial \mathrm{Y} < \partial C_v/\partial \mathrm{Y}$ for the artifact classes specified in Section 3.0.

Falsification: Verification cost per artifact declines faster than production cost per artifact as volume scales. Specifically: $\alpha \leq 1$ in the cost function $C_v(\mathrm{Y}) = \mathrm{c\_0Y} + \mathrm{c\_1}Y^{\alpha}$.

Test: Regress log(review_time) on log(artifact_volume) within organizations. If coefficient $\leq$ 1, non-linearity does not bind.

Data source: GitHub PR review times; internal code review logs with timestamps.

Limitation: Review time is a proxy for cost; cognitive effort is not directly observed.

F2. Cascade Formation (Proposition 1)

Core claim: When C_v - $C_f \geq B_n$et, agents forward regardless of private beliefs, producing information blockage.

Falsification: Agents continue to verify at stable rates even as (C_v - C_f) / $B_n$et increases. Forwarding behavior is independent of cost structure.

Test: Track verification intensity (review depth, revision cycles) as a function of artifact volume and time pressure. If verification intensity remains stable under high-volume conditions, the cost threshold is not binding.

Data source: PR review metadata; code review quality scores; audit coverage rates.

F3. Synthetic Productivity (Claim 1)

Core claim: Measured TFP diverges from verification-adjusted TFP as verification intensity declines.

Falsification: Correction burden (R) grows slower than or equal to throughput gains, and verification intensity remains stable.

Test: Compute wedge Δ = Y - (Y $\cdot$ v - R) over time. If Δ is stable or declining, synthetic productivity is not accumulating.

Data source: Throughput metrics (commits, tickets closed); remediation metrics (churn, reverts, incident response hours).

F4. Recognition Lag Extension (Claim 2)

Core claim: Increasing α_t (model-mediated substrate share) extends τ_d (detection lag).

Falsification: Detection latency for seeded defects remains constant as α_t rises.

Test: Inject known anomalies at varying α_t levels; measure time-to-detection. If τ_d is independent of α_t, substrate contamination does not affect detection.

Data source: Controlled experiments with seeded defects; incident detection timelines.

F5. Market Selection (Proposition 2)

Core claim: Low-verification organizations gain market share during stable periods.

Falsification: High-verification organizations maintain or increase share during stable periods, or low-verification organizations are systematically eliminated before crises.

Test: Track verification proxies (audit coverage, review depth) against market share or resource allocation. If high-θ proxies predict share gains in stable periods, selection dynamics are reversed.

Data source: Industry surveys; public company filings with quality metrics; startup failure rates by verification posture.

F6. Verification Capacity Erosion (H5)

Core claim: $\mathrm{d}\theta_t/\mathrm{dt} \leq 0$ and $\mathrm{d}\kappa_t/\mathrm{dt} < 0$ (capacity and capability decline under sustained underuse).

Falsification: Verification capacity and capability remain stable or improve despite reduced utilization.

Test: Track seeded defect detection rates over time. If detection rates remain stable even as review volume declines, capability is not eroding.

Data source: Controlled seeded-defect experiments; reviewer performance metrics.

Boundary Cases Where Predictions Weaken

The framework's predictions weaken or fail under the following conditions:

1. **Tight formal contracts.** When artifacts have unambiguous correctness criteria (types, proofs, model checking), automated verification can scale with production. The cost asymmetry does not bind.
2. **High-coverage automated tests.** When fast, comprehensive test suites provide immediate ground-truth feedback, verification cost may scale linearly or sub-linearly with output.
3. **Observable verification stock (θ).** If verification capacity can be audited, certified, or credibly disclosed, markets can price it pre-crisis and selection against verification weakens.
4. **Frequent crises.** If crises arrive frequently relative to stable periods (T_c large relative to T_s), crisis-period selection dominates and high-verification strategies are favored.
5. **Organizational learning.** If organizations can rapidly increase verification capacity when problems become visible, the cascade equilibrium is not absorbing.
6. **Resource constraints on generation.** If production costs rise (energy, compute, regulatory barriers), the cost asymmetry may reverse and verification becomes the cheaper margin.

Joint Prediction

If the framework is correct, we should observe the following pattern over time:

1. Throughput (Y) rises faster than verification intensity (v)
2. Correction burden (R) rises, but with lag
3. Detection latency (τ_d) increases as substrate share (α_t) rises
4. Variance in routine metrics declines (occurrence lag)
5. Severity of detected problems increases (compounding)
6. Market share shifts toward low-verification organizations during stable periods

If three or more of these predictions fail simultaneously, the framework's core mechanism is not operating as described.

---

12 Conclusion

The argument of this paper is simple to state and difficult to escape.

When producing an artifact becomes cheaper than verifying it, rational agents stop verifying. This is not a moral failure; it is an equilibrium. The agents are optimizing correctly given the cost structure they face. The pathology is structural.

From this single asymmetry, a cascade of consequences follows. Information stops aggregating. Measurement systems become self-referential. Market selection favors strategies that will fail

under stress. Organizations that invest in verification are outcompeted by organizations that invest in throughput.

None of this requires malice. None of this requires stupidity. The mechanism operates on rational actors making locally optimal decisions. That is what makes it difficult to reverse: there is no villain to remove, no error to correct. The error is the cost structure itself.

I have tried to be precise about what this framework does and does not claim. It does not claim that AI is bad, that productivity gains are illusory, or that all model-generated content is unreliable. It claims that a specific cost asymmetry—generation cheaper than verification—changes equilibrium behavior in ways that current measurement systems do not capture.

The empirical agenda is the honest part of this paper. I have specified conditions under which the framework would be falsified, and I have proposed instrumentation to test them. If verification intensity proves robust to volume increases, if cascade fragility does not scale with coupling, if recognition lags do not extend with substrate contamination—then the framework is wrong, and I will have learned something.

But if the framework is approximately right, we are in a regime where the signals of trouble are systematically suppressed by the same dynamics that create the trouble. The forcing event, when it comes, will be a surprise to everyone except those who were already looking at the right variables.

The variables to watch are not throughput, velocity, or artifacts shipped. They are verification depth, remediation burden, and epistemic debt.

The dashboard can stay green while the balance sheet compounds in the dark. This paper is an attempt to read the balance sheet before the audit.

Reality is not free. The bill comes due. The only question is whether we see it before it arrives.

12.1 Acknowledgments

The "Double Minsky" framework in this paper owes a debt to Marvin Minsky and Hyman Minsky. The parallel was first noticed, as many things are, in conversation with a cat named Marvin, who has since supervised most of my thinking about resilience, redundancy, and the limits of pattern completion. To be precise, credit belongs to three individual Minskys and two Marvins. To both Marvin Minskys (feline and human), thank you.

I have no institutional affiliations or grant funding to disclose. This is independent work, unconnected to other projects, reconciling grad school nostalgia with the realization that I can still research; it's just not due by midnight. This framework is provisional. These hypotheses are bets. I've tried to be clear about what I know and what I'm inferring; it's not always clear on the inside.

12.2 References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3), 488–500.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3), 797–817.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992–1026.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work. *NBER Working Paper No. 31161*.

GitClear (2024). *Coding on Copilot: 2023 Data Suggests Downward Pressure on Code Quality*. Technical Report. https://www.gitclear.com

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3), 393–408.

International Energy Agency (2024). *Electricity 2024: Analysis and Forecast to 2026*. IEA, Paris. https://www.iea.org

Martin, R. C. (2008). *Clean Code: A Handbook of Agile Software Craftsmanship*. Prentice Hall.

Minsky, H. P. (1986). *Stabilizing an Unstable Economy*. Yale University Press.

Minsky, M. (1986). *The Society of Mind*. Simon & Schuster.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192.

Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. *2022 IEEE Symposium on Security and Privacy (SP)*, 754–768.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590*.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press.