

Detecting Silent Governance Failure: Semantic Entropy, Dissent Exhaustion, and Circuit Breakers

Takashi Sato

Independent Researcher, Japan

Email: i@takashisato.me

ABSTRACT

Current AI governance focuses on the noise of failure—errors, bias, and misuse. However, in high-stakes bureaucracy, the most catastrophic failures manifest as silence. We introduce *Resolution Collapse*: a condition where human oversight persists as a procedural ritual, as organizational efficiency smooths away the friction required to distinguish routine cases from exceptions.

To address this, we shift the focus from error prevention to failure legibility. We propose *Governance Drift Indicators (GDIs)* to detect the semantic and temporal erosion of judgment before it becomes epistemically irreversible. Furthermore, we argue that when oversight collapses, "soft" interventions like retraining are insufficient. Instead, we introduce a *Circuit Breaker* architecture: a structural mechanism designed to halt or isolate AI-assisted processes when epistemic capacity is lost. Finally, we define the concept of a *Proper Ending*—a controlled termination that preserves institutional memory. These contributions reframe AI governance not as the pursuit of uninterrupted operation, but as the capacity to stop systems transparently and responsibly.

CCS Concepts

- Social and professional topics → Computing / technology policy; Technology audits;
- Human-centered computing → Collaborative and social computing systems and tools.

Keywords

AI Governance; Resolution Collapse; Circuit Breakers; Governance Drift; Semantic Entropy; Sociotechnical Failure; Structural Governance Failure; Proper Ending

This third part completes a three-part work on workflow-centric AI governance.

It does not propose a solution, but documents the conditions under which governance itself must stop.

1. INTRODUCTION

The most dangerous state of an AI-assisted bureaucracy is not conflict, but seamlessness. While traditional governance literature focuses on preventing visible malfunctions—such as algorithmic bias or erroneous outcomes—this focus overlooks a more fundamental pathology: the gradual erosion of human contestability. In high-stakes environments, systems often fail not by crashing, but by operating smoothly while human judgment quietly withdraws from the loop. We argue that this "silence"—where decisions are processed efficiently without meaningful scrutiny—represents the deepest form of governance failure, precisely because it renders itself **unobservable** to standard metrics.

Current approaches to this problem have focused on structural intervention. Notably, Part I of this three-part work [1] has proposed architectures designed to compel human deliberation through procedural friction, aiming to prevent the passive acceptance of AI outputs. These frameworks posit that introducing specific "governance gates" can mechanically sustain human engagement. However, subsequent theoretical analyses of organizational failure in Part II of this three-part work [2] suggest that even such carefully designed systems are susceptible to structural governance drift—a condition where mechanisms degrade into ritualized procedures under the thermodynamic pressures of efficiency and cognitive economy.

In this third part, we address a critical operational gap left by these architectural and theoretical frameworks: what should be done when governance failure itself becomes unobservable? When oversight persists only as a procedural ritual, conventional remedies such as retraining or ethical guidelines offer little traction. In these circumstances, the problem is no longer how to improve judgment, but how to recognize that judgment has already ceased to function. To address this problem, we shift the focus of AI governance from error

prevention to failure legibility. We introduce the concept of Resolution Collapse, a condition in which human decision-makers lose the granularity required to perceive and contest failures.

Building on this concept, we propose *Governance Drift Indicators (GDIs)* as operational signals for detecting the onset of such collapse, and argue for the necessity of structural intervention mechanisms—Circuit Breakers—that can suspend AI-assisted decision-making once human oversight has crossed an *irreversible threshold*.

2. RESOLUTION COLLAPSE AS A GOVERNANCE FAILURE

Governance failures in AI-assisted decision-making are often framed as problems of incorrect outcomes or biased judgments. This framing, however, presupposes that human operators retain the capacity to recognize and contest such failures. In practice, a more fundamental breakdown occurs when this capacity itself erodes. We conceptualize that breakdown as *Resolution Collapse*.

2.1 Defining Resolution

To understand this failure, we must first define the specific quality of oversight that degrades. **We define Resolution** as the granularity with which human decision-makers perceive, interpret, and evaluate individual cases within a governed process. **High-resolution judgment** allows operators to recognize contextual nuance, identify exceptions, and articulate reasons for disagreement. **Low-resolution judgment**, by contrast, reduces cases to abstract categories or patterns, suppressing contextual detail in favor of procedural consistency.

2.2 The Mechanism of Collapse

Importantly, *resolution collapse* should not be understood as a result of negligence, incompetence, or malicious intent. **Rather, it emerges primarily** as an adaptive response to structural pressures within bureaucratic and organizational environments. This erosion aligns with what has been described in Part II of this three-part work as structural or "pre-abuse" governance failure [2], where structural pressures force operators into a state of cognitive externalization—validating procedure rather than substance. Under sustained demands for efficiency, consistency, and risk avoidance, human operators are incentivized to

compress complex judgments into simplified representations that minimize cognitive effort and personal liability. **Among these pressures, fear plays a central role.** When accountability mechanisms punish deviation more readily than conformity, human decision-makers learn that resisting automated recommendations carries asymmetric risk. At the same time, institutional conceptions of fairness often equate justice with uniform treatment, encouraging the suppression of case-specific distinctions. Together, these forces reward low-resolution judgment as a **rational survival strategy** within the system.

2.3 The Stabilization of Degraded Oversight

As resolution declines, human oversight does not disappear abruptly. Instead, it persists in a degraded form. Reviews continue to be performed, explanations continue to be recorded, and approval workflows continue to function. Yet these activities increasingly lose their substantive connection to the underlying decision context. Oversight becomes procedural rather than deliberative, preserving the appearance of control while abandoning its function. This distinction is critical. Temporary reductions in resolution—such as those caused by workload spikes—do not necessarily constitute governance failure. *Resolution collapse*, as defined here, refers specifically to a condition in which the system no longer supports the recovery of high-resolution judgment. Once operators cease to perceive meaningful distinctions between cases, additional information, training, or ethical guidance fails to restore deliberative capacity.

3. GOVERNANCE DRIFT INDICATORS (GDIs): MEASURING THE INVISIBLE

Governance failures associated with resolution collapse pose a fundamental measurement problem. By definition, resolution collapse renders failures increasingly unobservable to human operators. Traditional governance metrics—such as error rates or compliance scores—presuppose the continued presence of meaningful human judgment. When that judgment degrades, these metrics can remain stable or even improve, masking the underlying failure. To address this problem, we introduce *Governance Drift Indicators (GDIs)*: a set of operational signals designed to detect the erosion of human judgment before governance failure becomes irreversible.

These indicators are intended as diagnostic signals rather than prescriptive evaluative metrics and require contextual interpretation.

Crucially, the GDI framework does not attempt to infer intent or ethics. Instead, it treats governance degradation as a systemic process reflected in observable behavioral and informational patterns.

3.1 Semantic Entropy Decay

The first indicator concerns the informational content of human-generated explanations. In high-resolution governance, human operators articulate reasons for decisions using diverse language that reflects case-specific considerations. Such explanations exhibit high semantic variability. As resolution collapses, linguistic diversity declines. Explanatory text converges toward repetitive phrases, boilerplate justifications, and formulaic references to guidelines (e.g., "Approved per standard procedure"). We characterize this process as Semantic Entropy Decay: a reduction in the informational richness of language used to justify decisions. This phenomenon reflects a form of "responsibility inversion" discussed in the sociotechnical analyses in Part II of this three-part work [2], where documentation serves liability rather than epistemic verification.

$$\text{SED}(t) \propto -\frac{d}{dt} \left(\frac{V(t)}{L(t)} \right)$$

Where $V(t)$ is the cumulative vocabulary count and $L(t)$ is the

Operationally, this may be approximated through measures of lexical diversity or compression ratios in free-text fields. **A sustained decline constitutes a strong signal** that human deliberation is being replaced by procedural conformity. In contexts using generative AI, the uncritical copying of LLM-generated explanations—even if lexically diverse—should also be treated as a form of entropy decay, as it represents the displacement of human judgment.

3.2 Temporal Compression

The second indicator concerns time. Meaningful judgment requires temporal space for consideration. In well-functioning systems, decision times vary, reflecting differences in case complexity. **Temporal Compression** refers to a pattern in which the duration between AI output presentation and human decision shrinks toward a narrow, uniform range approaching the physiological lower bounds of cognitive processing (e.g., *basic motor response time*).

$$\text{TC}(t) \propto -\frac{d}{dt} (\sigma^2(\Delta t_i))$$

Where $\sigma^2(\Delta t_i)$ is the variance of decision processing times across cases i .

The most salient signal is not merely shorter average times, but the disappearance of variance. When complex and routine cases are processed at the same speed, judgment has effectively been reduced to execution.

3.3 Exhaustion of Dissent

The third indicator captures the erosion of resistance. In high-resolution systems, **meaningful disagreement between human judgment and AI recommendations should manifest** as rejections, modifications, or escalations. As resolution collapses, dissent becomes increasingly rare. Operators learn that challenging automated outputs carries personal risk while offering little reward. Over time, rejection rates decline and approval becomes the default. This phenomenon is the **Exhaustion of Dissent**. When dissent approaches zero independently of model performance improvements, governance has shifted from deliberation to acquiescence.

These indicators are intentionally framed as governance diagnostics rather than performance metrics, and are designed to trigger institutional reflection and containment, not automated optimization.

Box 1: Minimal Operationalization of Governance Drift Indicators (Illustrative)

To clarify the operational intent of Governance Drift Indicators (GDIs) and to establish a point of falsifiability, we provide a minimal measurement protocol. This specification is illustrative only and has not yet been empirically validated. It is intended not as a definitive implementation guide, but as a conceptual scaffolding for future empirical validation.

GDI	Measurement Proxy	Warning Signal
Semantic Entropy Decay (SED)	Lexical diversity (e.g., Type-Token Ratio or n-gram entropy) computed over	A sustained decline in the diversity metric over k consecutive

	Tw, a sliding window of the last N human-authored decision explanations or review comments.	windows, indicating convergence toward a minimal vocabulary.
Temporal Compression (TC)	Variance of decision time ($\sigma^2\Delta t$) across cases of heterogeneous complexity.	Collapse of variance ($\sigma^2\Delta t \rightarrow \epsilon$), suggesting review time is no longer proportional to case complexity.
Exhaustion of Dissent (ED)	Dissent rate defined as the ratio of human overrides, modifications, or escalations relative to automated recommendations.	A monotonic decline in dissent independent of verified model performance improvements.

Irreversible Threshold. The threshold is never defined by a single metric. It is crossed when multiple indicators (e.g., two out of three) simultaneously exceed predefined, context-specific limits for a sustained duration, necessitating non-negotiable containment.

3.4 Case Vignette: Resolution Collapse in Retrospect (The Dutch Childcare Benefits Scandal)

To illustrate how *Resolution Collapse* manifests in practice and how *Governance Drift Indicators (GDIs)* become observable in retrospect, we examine the Dutch Childcare Benefits Scandal (2013–2020). This vignette is **not intended as a political or legal assessment** of responsibility, but as an analytic reinterpretation of a widely investigated case.

In this episode, an automated enforcement system systematically treated minor administrative discrepancies (e.g., missing signatures or delayed documentation) as indicators of intentional fraud. Under sustained political and organizational pressure for efficiency, the bureaucracy collapsed the critical distinction between routine error and deliberate misconduct. Human caseworkers formally remained in

the loop, yet their role was reduced to procedural execution, rendering meaningful judgment epistemically unavailable.

Several GDIs were concurrently present. First, *Semantic Entropy Decay* was evident in official correspondence. Archived rejection and recoupment notices were highly uniform, relying on rigid boilerplate language that failed to engage with individual circumstances. Retrospective analysis using *lexical diversity measures* could have revealed near-zero semantic variability across cases, indicating the erosion of case-specific reasoning.

Second, *Exhaustion of Dissent* was observable within the organization. Internal warnings from legal advisors, ombudsmen, and frontline staff were documented, yet systematically neutralized. *Log-based reconstruction* of escalation records could have shown that while the frequency of dissent attempts remained stable or even increased, their practical efficacy—the rate at which objections altered outcomes—approached zero.

Importantly, these indicators could have been reconstructed retrospectively using artifacts that were already available at the time, including decision timestamps, standardized correspondence templates, and internal records of objections and legal warnings. The absence of real-time monitoring was therefore not a limitation of measurability, but a **governance choice that left these signals unobserved**.

This sustained coupling of *semantic entropy decay* and dissent exhaustion distinguishes genuine resolution collapse from routine bureaucratic efficiency. Had such coupled signals been monitored, a structural intervention—such as redirecting enforcement into a mandatory human audit mode—could have occurred before epistemic capacity became irreversibly compromised.

Beyond traditional state bureaucracy, recent incidents suggest that Resolution Collapse can occur acutely within high-stakes knowledge work reliant on expert judgment, such as specialized consulting and legal compliance. In settings where generative AI tools have been broadly adopted, recent public reports indicate that expert-authored documents, despite undergoing formal internal review, have contained fabricated

foundational evidence, including non-existent citations and legal precedents. This acute form of collapse confirms that oversight persists merely as a procedural ritual—the critical friction required for non-routine verification is effectively smoothed away—even in domains defined by the complexity and quality of their evidence base.¹

4. THE IRREVERSIBLE THRESHOLD

The indicators introduced above raise a critical question: at what point does governance degradation cease to be recoverable? Not all instances of governance drift warrant structural intervention. This section distinguishes reversible degradation from a qualitatively different condition: **the Irreversible Threshold**. This phenomenon can be understood through the lens of hysteresis in control systems. Once a critical boundary is crossed, simply reducing workload or providing more information is unlikely to restore high-resolution judgment. **The system has entered a state in which** judgment loss becomes structurally stabilized. While the simulation studies on compliance decay in Part I of this three-part work [1] suggest that oversight efficacy exhibits a non-linear phase transition, operationalizing this threshold requires observable metrics. The *irreversible threshold* is identified by the **persistence and coupling** of multiple GDI signals. To distinguish genuine collapse from the efficient heuristics of expert operators, this threshold is never defined by a single metric. Instead, it requires the sustained and simultaneous manifestation of multiple signals—such as semantic entropy decay occurring alongside the exhaustion of dissent—over a prolonged period. When semantic entropy remains low despite policy changes, and dissent fails to re-emerge despite encouragement, the system has lost the capacity to know whether its decisions are correct. Attempting to restore judgment beyond this threshold through "soft" interventions (training, reminders) is counterproductive. The recognition of an *irreversible threshold* necessitates a shift from correction to containment. **Governance failure, in this sense, is defined by the loss of observability.**

¹ This phenomenon is exemplified by recent public reports involving major consulting audits for public sector clients, where multiple erroneous citations generated by AI were detected, ultimately leading to contract disputes. The incident underscores the rapid erosion of epistemic capacity when professional

5. CIRCUIT BREAKER ARCHITECTURE

Once the irreversible threshold has been crossed, governance intervention must shift toward structural containment. This section introduces the *Circuit Breaker* as an architectural mechanism designed to halt or isolate AI-assisted decision processes when governance drift becomes unrecoverable.

5.1 Design Principles

The *circuit breaker* concept is borrowed from electrical engineering and financial markets. Its objective is not to fix erroneous decisions, but to prevent continued operation under conditions of epistemic blindness.

Non-negotiability: Activation must not be subject to override or discretionary delay. The purpose is to remove decision authority when decision capacity is compromised.

Proportional Isolation: Isolation may range from suppressing AI outputs (forcing manual review) to suspending the workflow entirely.

Reversibility with Accountability: The system preserves logs and states necessary for post-hoc analysis.

Circuit breakers must distinguish between **fail-safe** and **fail-open** contexts. In domains where a single erroneous action is catastrophic (e.g., weapons systems or high-frequency trading), a **fail-safe** strategy involving complete lockout is appropriate. By contrast, in essential public services such as welfare distribution, healthcare eligibility, or migration processing, suspension should default to a **fail-open** posture: cases are redirected to mandatory human review queues or provisionally granted, rather than denied outright. Under conditions of *resolution collapse*, automated denial constitutes a distinct and immediate form of harm, making **fail-open redirection** the ethically preferable safeguard. *Non-negotiability* in such settings requires both *technical enforcement* (e.g., interface constraints and log authority) and *institutional enforcement* (e.g., regulatory mandate or

verification becomes structurally dependent on automated outputs.

contractual obligation) to resist managerial or political override.

We acknowledge that activating a *circuit breaker* incurs significant operational costs and service interruptions. However, we argue that the long-term societal cost of maintaining a formally compliant but epistemically hollow bureaucracy—operating without meaningful human oversight—is **fundamentally higher and more dangerous**.

5.2 Triggering and Intervention

Circuit breakers are triggered by sustained patterns in GDIs—specifically, when multiple indicators simultaneously exceed predefined thresholds. Ideally, these breakers engage directly with workflow gates, such as those in the gate-based governance framework proposed in Part I of this three-part work [1], forcing a 'hard stop' at the interface level.

Mode A (Blind Mode): Suppress AI recommendations from the user interface. Human operators are forced to proceed without automated guidance, immediately restoring cognitive engagement (or halting if they cannot). If the removal of AI guidance causes the workflow to stall entirely, this failure is not a bug but a diagnostic result: it proves that the system has already succumbed to total cognitive dependency.

Mode B (Lockout): Disable automated approval pathways and redirect all cases to a separate audit queue.

Mode C (Suspension): Complete halt of the decision pipeline.

6. DESIGNING THE PROPER ENDING

The introduction of circuit breakers raises the question of what constitutes a *proper ending* for an AI-assisted governance system. In many contexts, termination is treated as failure. We argue the opposite: in high-stakes governance, the ability to terminate a system responsibly is a core requirement of legitimacy. A *Proper Ending* is defined as the controlled suspension of a decision process in a manner that preserves accountability, traceability, and institutional memory.

Traceability: The system records why it stopped (e.g., "Semantic entropy fell below 0.2").

Accountability: Responsibility is not displaced onto "technical error" but returned to the institution.

Contestability: The consequences of the termination remain visible to affected stakeholders.

Designing for proper endings reconfigures success. Rather than defining success as uninterrupted operation, systems are evaluated by their capacity to recognize when continued operation is no longer justifiable. Termination is not the negation of governance, but its final expression. **Crucially, proper endings are not irreversible annihilations. However, any reactivation after termination must be treated as a new governance decision, not as a continuation of the prior system.** The burden of justification shifts from explaining why the system should stop to explaining why it should run again.

7. CONCLUSION

We have argued that the central challenge of AI governance is not the prevention of failure, but the management of failure's visibility. We conceptualized the erosion of oversight as Resolution Collapse and introduced Governance Drift Indicators (GDIs) to detect it. We proposed Circuit Breakers as a structural mechanism for halting AI-assisted processes once epistemic collapse makes them unmanageable. Finally, we articulated the concept of a Proper Ending. **Allowing systems to end is not pessimism—it is realism.** Governance that cannot acknowledge its own limits inevitably exceeds them. In this sense, the legitimacy of AI governance should be judged not by how seamlessly systems operate, but by whether institutions retain the capacity to recognize, justify, and enact a deliberate stopping point when judgment itself has been exhausted.

Limitations and Operational Readiness

Finally, we emphasize that the Governance Drift Indicators (GDIs) proposed here are designed as diagnostic signals, not evaluative verdicts. We do not claim that these metrics are currently ready for automated deployment; rather, we argue that the conceptual scaffolding for detecting resolution collapse must precede its measurement. The absence of large-scale empirical validation in this work is not merely a limitation of data availability, but a reflection of the core problem: the structural unobservability of

the failure itself. Moving from conceptual definition to empirical monitoring is not solely a technical challenge but a governance decision—one that requires institutions to first acknowledge that their silence may be a sound.

This absence of large-scale empirical validation is not an oversight, but a deliberate methodological boundary. The purpose of this work is not to optimize governance metrics, but to formally define the conditions under which governance itself becomes epistemically unobservable, at which point conventional empirical validation loses meaning.

The absence of large-scale empirical validation should not be interpreted solely as a methodological limitation. Resolution collapse, by definition, manifests as a loss of observability within decision-making systems, making such failures difficult to detect using standard performance metrics. Nevertheless, the indicators proposed here remain falsifiable through retrospective reconstruction using archival decision logs, timestamps, and explanatory artifacts, as illustrated by the Dutch Childcare Benefits case.

8. REFERENCES

- [1] Sato, T. (2025). Workflow-Centric AI Governance: A Sociotechnical Architecture for Accountable Human–AI Decision Systems. Part I of a three-part work on workflow-centric AI governance.
- [2] Sato, T. (2025). Forecasting Failure: Why Structurally Sound AI Governance Collapses Before It Is Abused. Part II of a three-part work on workflow-centric AI governance.
- [3] Power, M. (1997). *The Audit Society: Rituals of Verification*. Oxford University Press.
- [4] Almada, M. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems. *Proceedings of the 17th International Conference on Artificial Intelligence and Law (ICAIL '19)*.
- [5] Elish, M.C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*. 5, 40–60.
- [6] Parasuraman, R. and Manzey, D.H. (2010). Complacency and Bias in Human Interaction with Automation: A Taxonomy and Review. *Human Factors*. 52, 3, 381–410.
- [7] Raji, I.D. et al. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT '20)*.
- [8] Scott, J.C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press.
- [9] Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. University of Chicago Press.