Semantic Commitment as a Structural Constraint in Long-Horizon Adaptive Systems

Maksim Barziankou, Poznan 2026

Abstract

Adaptive systems deployed over long horizons increasingly operate in environments where decisions carry irreversible structural consequences. In such settings, failure may arise even when observations are correct, behavior is locally admissible, and no adversarial interference is present. This work identifies a distinct and underexamined failure mode in which transient contextual salience induces premature semantic commitment, allowing unstable interpretations to authorize irreversible structural operations.

We argue that this class of failure cannot be eliminated through improved perception accuracy, probabilistic inference, or reward optimization. Instead, it exposes a missing architectural layer governing when meaning is permitted to act. The central claim is that semantic commitment is not an inferential outcome but a structural authorization, and that long-horizon coherence depends on regulating this authorization independently of instantaneous confidence or performance.

We outline an architectural framework that separates semantic interpretation from semantic commitment, constrains commitment through an internal temporal admissibility process, and treats organizational regime transitions as protected structural operations with irreducible cost. The framework is presented at the level of architectural principles and invariants, without reliance on specific algorithms or optimization schemes, and is intended to serve as prior art for adaptive systems requiring semantic safety under contextual uncertainty.

By reframing semantic failure as a problem of authorization rather than accuracy, this work opens a new perspective on long-horizon adaptation—one in which the discipline of restraint, rather than the pursuit of ever-faster inference, determines whether meaning becomes a source of coherence or collapse.

Introduction

Adaptive systems are increasingly deployed in environments characterized by uncertainty, partial observability, delayed feedback, and irreversible consequences. In such settings, the primary risk is not merely suboptimal action, but structurally consequential action: an intervention that alters the system's organizational state, consumes irreducible structural cost, or degrades viability and identity in ways that cannot be recovered through subsequent local correction. These environments amplify a gap between short-horizon performance and long-horizon survival: a system may remain locally correct while drifting toward regime instability, semantic brittleness, or irreversible collapse.

A common assumption in contemporary design is that interpretive failure is primarily a deficiency of perception accuracy, probabilistic inference, learning capacity, or optimization quality. Under this view, increasing sensor fidelity, model expressivity, or objective sophistication should progressively reduce incorrect decisions. However, long-horizon failures persist in a distinct class of situations where observations are correct, behavior is locally admissible, and no adversarial manipulation is present, yet the system still commits to an incorrect meaning that becomes structurally actionable. In these cases, the failure is not attributable to erroneous data or insufficient optimization, but to the timing and admissibility of semantic commitment itself.

This work addresses a failure mode we refer to as *context capture*. Context capture occurs when a transient contextual anchor—often induced by salience, coincidence, partial history, or ambiguous coupling—forces premature semantic closure. The resulting interpretation may be internally coherent and locally plausible, yet structurally inadmissible as a basis for irreversible operations. The critical observation is that semantic error can emerge even under correct sensing and admissible local control, because the system binds observations into a coherent explanatory structure without sufficient temporal stabilization, and then allows that structure to authorize irreversible change. In other words, the system is not wrong about what it sees, but wrong about what it permits meaning to do.

We therefore distinguish two architectural roles that are frequently conflated: *semantic interpretation* and *semantic commitment*. Interpretation is the generation of explanatory hypotheses from interaction-derived signals; it must remain exploratory, abundant, and reversible under uncertainty. Commitment is the authorization for interpretations to influence irreversible structural operations such as transitions between organizational regimes, structural reconfiguration, or identity-affecting interventions. Commitment is not an inferential act; it is a structural act. When these roles are not separated, transient interpretations become a hidden decision surface, enabling context-level plausibility to trigger regime-level consequences.

The central claim of this paper is that long-horizon coherence requires treating semantic commitment as a governed authorization process rather than a byproduct of inference. In particular, the admissibility of commitment must be regulated by architectural constraints that are independent of instantaneous confidence, probability, or reward signals. The core principle is simple: a system may generate hypotheses freely, but it must not allow those hypotheses to act irreversibly until meaning has stabilized under an appropriate temporal notion of admissibility. This paper adopts an internal-time perspective for that stabilization notion: meaning is considered admissible for commitment only when it persists under an internal temporal measure that reflects uncertainty, reliability degradation, and coupling conditions. The exact estimators and mappings are not required for the argument and are intentionally left non-operational in this prior-art presentation.

Our objective is not to propose a learning algorithm, a control law, or a new optimization objective. We do not assume perfect perception, exhaustive world models, or richer reward functions. Instead, we articulate a missing architectural layer governing *when* meaning is permitted to act, how irreversible operations are protected from transient contextual interpretations, and why viability and identity preservation must be enforced through constraint rather than performance maximization. The contribution is therefore architectural: a framing of semantic failure as an authorization problem, together with the invariants required to prevent context capture from becoming structurally consequential.

The remainder of the paper proceeds as follows. We first formalize context capture as a failure mode that can occur without perceptual error and without adversarial interference. We then articulate the separation between interpretation and commitment, and motivate internal time as the appropriate regulator of semantic admissibility under uncertainty. Next, we discuss organizational regimes and structural cost to clarify why regime transitions constitute irreversible structural operations that must be protected from context-level plausibility. Finally, we position the framework relative to existing approaches and summarize the invariants that enable interpretive richness without structural vulnerability under long-horizon operation.

Context Capture Without Error

We define *context capture* as a failure mode in which transient contextual salience induces premature semantic closure, binding unrelated or weakly coupled observations into a coherent explanatory structure. The defining feature of context capture is not the presence of incorrect data, but the improper elevation of a transient interpretive scaffold into an actionable semantic commitment.

Crucially, context capture may occur under conditions where observations are fully correct, no adversarial manipulation is present, and behavior remains locally admissible with respect to immediate objectives or constraints. In such cases, the system's perceptual and control subsystems may function nominally, and short-term performance metrics may indicate successful operation. Nevertheless, the system commits to a semantic interpretation that is structurally inadmissible when evaluated over a longer horizon.

The origin of the failure lies in the interaction between salience, partial observation, and incomplete causal history. A salient contextual anchor—such as temporal coincidence, peripheral correlation, or sudden novelty—can exert disproportionate interpretive pressure, forcing the system to resolve ambiguity prematurely. The resulting semantic closure may be internally coherent and locally plausible, yet unsupported by sufficient temporal stabilization. When such closure is permitted to authorize irreversible structural operations, semantic error becomes structurally consequential.

Context capture therefore differs fundamentally from classical perceptual error, statistical overfitting, or reward hacking. In perceptual error, the data are incorrect; in overfitting, the model generalizes poorly; in reward hacking, the objective is misaligned. In context capture, none of these conditions need to hold. The failure arises instead from incorrect authorization of meaning: the system allows a transient context-level interpretation to act as if it were stabilized meaning. This distinction motivates the need for an architectural mechanism that governs when semantic interpretations may become structurally actionable, independently of their local plausibility or confidence.

—

Semantic Interpretation vs. Semantic Commitment

We introduce a fundamental architectural distinction between *semantic interpretation* and *semantic commitment*. This distinction is essential for understanding how adaptive systems can remain coherent under uncertainty without suppressing interpretive richness.

Semantic interpretation refers to the generation of explanatory hypotheses or meanings from interaction-derived observations. Interpretation is an epistemic process: it explores possible structures that could account for observed signals and contextual cues. Under uncertainty, interpretation must remain unrestricted, abundant, and reversible. Premature restriction of interpretation reduces adaptability and leads to brittleness, particularly in environments with partial observability or delayed feedback.

Semantic commitment, by contrast, refers to the authorization for semantic interpretations to influence irreversible structural operations. Such operations include, but are not limited to, transitions between organizational regimes, structural reconfiguration, and identity-affecting interventions. Commitment is therefore not an epistemic act but a structural one. It determines when meaning is permitted to alter the system's organizational state in ways that cannot be trivially undone.

Conflating interpretation with commitment introduces a hidden decision surface within the interpretive process itself. When semantic interpretations are allowed to directly trigger irreversible operations, transient or unstable meanings may acquire structural authority simply by virtue of local coherence or salience. This coupling effectively bypasses any temporal stabilization requirement and exposes the system to context-driven structural failure.

By separating interpretation from commitment, an adaptive system can preserve exploratory inference while constraining structural risk. Interpretations may form, compete, and evolve freely, while commitment is governed by distinct admissibility criteria that account for temporal persistence, uncertainty, and long-horizon viability. This separation reframes semantic safety as an architectural authorization problem rather than a matter of improving inference accuracy or optimization performance.

Internal Time as a Regulator of Semantic Admissibility

We argue that semantic commitment cannot be governed by instantaneous confidence, probability, or reward signals. Such signals, while useful for local inference or short-horizon optimization, are inherently myopic: they reflect momentary estimates rather than the durability of meaning under uncertainty. Relying on these quantities to authorize irreversible structural operations conflates local plausibility with long-horizon admissibility.

Instead, semantic commitment must be regulated by an internal temporal measure that reflects reliability degradation, ambiguity, and coupling conditions. Internal time is distinct from wall-clock time and is not required to advance uniformly. Its role is not to measure duration, but to encode the admissibility of structural commitment under uncertain interaction. When reliability is high and coupling is stable, internal time may advance rapidly; when uncertainty increases, causal history is incomplete, or coupling degrades, internal time slows.

By governing admissibility rather than inference, internal time determines *when* a semantic interpretation may be permitted to act, not *which* interpretation is correct. Under degraded conditions, the slowing of internal time structurally delays commitment even when interpretations appear locally coherent or confident. This delay is not a heuristic hesitation but an architectural constraint that prevents transient contextual interpretations from becoming structurally actionable. In this sense, internal time functions as a stabilizing buffer between interpretation and commitment, ensuring that meaning must persist across admissible temporal structure before acquiring structural authority.

—

Organizational Regimes and Structural Cost

Adaptive systems operate within organizational regimes that define sustained modes of operation, characterized by differing intervention densities, coordination patterns, and structural costs. A regime is not a single action or control signal, but an organizational configuration that shapes how the system interacts with its environment over extended periods.

Transitions between organizational regimes are non-commutative and incur irreducible structural cost. Such costs cannot be fully recovered through subsequent local optimization, as regime transitions alter the system's internal organization, coupling structure, and future admissible actions. Consequently, regime transitions constitute irreversible structural operations in the architectural sense, even when they are locally beneficial or reversible in appearance.

Because of their structural impact, regime transitions must be protected from context-driven semantic interpretations. Allowing context-level meaning—formed under transient salience or partial observation—to trigger regime transitions exposes the system to unnecessary structural load and long-horizon degradation. Even when individual transitions appear justified in isolation, repeated context-driven switching accumulates structural cost, erodes viability, and destabilizes identity.

Protecting regime transitions therefore requires that semantic commitment be stabilized prior to authorization. Only meanings that have persisted under appropriate temporal admissibility should be allowed to induce regime-level change. This framing shifts the focus from selecting the "best" action to governing the conditions under which organizational change is permitted, reinforcing the need for architectural separation between interpretation, commitment, and structural consequence.

Viability Beyond Performance

Performance metrics may remain stable while system viability degrades invisibly. Short-horizon measures such as reward accumulation, accuracy, or immediate task success capture only local correctness; they do not reflect whether the system remains structurally capable of sustaining coherent operation over extended horizons. A system may therefore appear successful while progressively exhausting its admissible space of future actions.

Viability is a regime-level, history-dependent property. It reflects whether the system can continue to operate without incurring irreversible structural damage, identity loss, or collapse into unstable oscillation. Because viability depends on accumulated structural cost, past regime transitions, and unresolved commitments, it cannot be reduced to instantaneous performance indicators or optimized directly without distortion.

Semantic commitment must therefore be evaluated with respect to viability preservation rather than performance optimization. An interpretation that improves short-term performance may still be structurally inadmissible if it induces excessive regime switching, accumulates irreversible cost, or destabilizes identity. From this perspective, semantic safety is not a matter of choosing the most rewarding action, but of constraining which meanings are allowed to authorize structural change. Viability is preserved not through maximizing an objective, but through enforcing architectural limits on commitment under uncertainty.

—

Structural Self-Forgiveness and Structural Debt

Long-horizon adaptive systems inevitably accumulate structural self-penalty arising from reversible errors, aborted commitments, or transitions that were initiated but later deemed inadmissible. Even when such events do not result in irreversible damage, they may introduce inhibitory constraints that persist within the system's organizational structure. Over time, the accumulation of these constraints reduces the admissible space of future actions and regime transitions.

We refer to this accumulation as *structural debt*. Structural debt is not synonymous with error or failure; it is the residual cost of cautious or corrective behavior under uncertainty. When left unregulated, structural debt progressively narrows the system's operational flexibility, leading either to paralysis, in which no commitment is deemed admissible, or to over-regulation, in which compensatory interventions become unstable and excessive.

To prevent such outcomes, we introduce *Structural Self-Forgiveness* as a controlled architectural operation that discharges structural self-penalty after events have been structurally closed. Structural closure denotes the point at which an error, aborted commitment, or non-admissible transition has been fully accounted for and no longer requires active inhibition. Structural Self-Forgiveness does not erase memory, meaning, or historical cost; nor does it legitimize inadmissible actions retroactively. Instead, it releases inhibitory constraints that are no longer required for safety, restoring admissible flexibility without compromising accountability.

Structural Self-Forgiveness is therefore neither forgetting nor tolerance of damage. It is a necessary complement to commitment gating in long-horizon systems, ensuring that caution does not itself become a source of structural failure. By enabling controlled discharge of structural debt after closure, the system preserves adaptability and avoids rigidity or paralysis while maintaining protection against irreversible harm.

Architectural Invariants

The proposed framework is governed by a set of architectural invariants that constrain how meaning may influence structure in long-horizon adaptive systems. These invariants are not implementation-specific rules, but necessary conditions for preserving coherence, viability, and identity under uncertainty.

Semantic interpretation must remain unrestricted. The system must be free to generate, explore, and revise explanatory hypotheses without penalty or premature pruning. Restricting interpretation under uncertainty reduces adaptability and shifts error upstream, increasing brittleness rather than safety.

Semantic commitment must be gated and reversible until stabilized. Authorization for semantic interpretations to influence irreversible structural operations must be explicitly separated from interpretation itself. Until stabilization criteria are satisfied, commitments must remain provisional and retractable, regardless of local plausibility.

Meaning stabilization is governed by internal time, not confidence, probability, or reward. Instantaneous metrics capture momentary belief strength but do not encode durability under uncertainty. Admissibility of commitment therefore depends on persistence across an internal temporal measure that reflects reliability, ambiguity, and coupling conditions.

Regime transitions incur irreducible structural cost. Transitions between organizational regimes alter the system's internal organization in ways that cannot be fully recovered through subsequent local optimization. Such transitions are therefore treated as irreversible structural operations requiring protection from transient semantic influence.

Structural geometry used for stabilization must not be exposed to the primary agent. Any metrics, criteria, or internal representations used to govern stabilization and admissibility must remain non-reconstructible from the agent's perspective. Exposing such geometry enables second-order optimization and undermines the integrity of commitment gating.

Viability and identity are preserved through constraint, not optimization. Long-horizon coherence cannot be achieved by directly optimizing viability or identity metrics without distortion. Instead, these properties are preserved indirectly through architectural constraints that limit when and how meaning is permitted to act.

Together, these invariants define an architectural boundary within which interpretive richness can coexist with structural safety.

—

Positioning with Respect to Prior Art

This work does not propose a learning algorithm, a control policy, or a new optimization objective. It does not rely on improved perception accuracy, richer reward functions, or exhaustive world models to address long-horizon failure. Instead, it introduces an architectural constraint layer governing semantic authorization under contextual uncertainty.

Prior approaches typically address semantic error by refining inference, adjusting objectives, or increasing model capacity. In contrast, the present framework treats semantic failure as an authorization problem rather than an inferential one. The contribution lies in separating interpretation from commitment, regulating commitment through internal time rather than confidence or reward, and protecting irreversible structural operations from transient contextual influence.

The framework is compatible with a wide range of adaptive systems, including artificial agents, robotic platforms, distributed decision systems, and cyber-physical systems. It is agnostic to specific learning paradigms, control laws, or representations, and may be integrated as an architectural layer atop existing systems. As such, it is intended to complement, rather than replace, existing approaches by addressing a failure mode that arises independently of perceptual accuracy or optimization quality.

Conclusion

Long-horizon failure in adaptive systems is not fundamentally a problem of intelligence, accuracy, or optimization. Systems fail not because they perceive the world incorrectly, nor because they lack sufficient computational power or expressive models, but because they permit meaning to act before it is structurally admissible. The decisive error lies not in what the system infers, but in when it allows inference to become irreversible.

This work has argued that semantic safety is an architectural problem. When semantic interpretation and semantic commitment are conflated, transient context, salience, or coincidence can silently acquire structural authority. Under such conditions, even correct observations and locally admissible behavior may culminate in irreversible damage. No increase in confidence calibration, probabilistic sophistication, or reward shaping can resolve this failure mode, because the vulnerability is not epistemic but structural.

By separating interpretation from commitment, adaptive systems can remain open to rich, exploratory meaning formation without exposing themselves to context-driven collapse. By regulating commitment through internal time rather than instantaneous belief or reward, systems gain a principled mechanism for delaying irreversible action under uncertainty. By treating regime transitions as protected structural operations with irreducible cost, systems acknowledge that some decisions change not only what they do, but what they are. And by preserving viability and identity through constraint rather than optimization, systems avoid the paradox of attempting to maximize the very properties that must instead be safeguarded.

The result is a reframing of adaptation itself. Intelligence is no longer defined solely by the ability to infer or optimize, but by the discipline with which meaning is allowed to act. In long-horizon environments, survival depends not on choosing the right interpretation quickly, but on refusing to let the wrong interpretation act too soon. This architectural discipline does not suppress intelligence; it enables it to persist.

In this sense, the central lesson is stark and unavoidable: long-horizon coherence is not achieved by making systems smarter, faster, or more confident. It is achieved by teaching systems restraint—by embedding, at the architectural level, the principle that meaning must earn the right to act.

MxBv, 2025 - 2026

DOI: https://doi.org/10.5281/zenodo.18212229