Non-Causal Observation of Viability and Identity in Long-Horizon Adaptive Systems

Maksim Barziankou, Poznań 2026
Project Petronus research

Abstract

Long-horizon adaptive systems frequently fail not due to incorrect actions or local performance errors, but due to gradual loss of structural viability and erosion of organizational identity. Existing control and learning frameworks predominantly regulate behavior through causal mechanisms embedded in action selection, optimization, or policy adaptation. This work identifies a fundamental limitation of such approaches: the absence of an architectural perspective that allows viability and identity continuity to be observed without becoming causally operative.

We argue that long-horizon viability and identity are global, temporally extended properties that cannot be reliably assessed from within an agent's decision-making process without distorting behavior. When evaluation of such properties is made causal, adaptive systems tend to optimize against their own monitoring signals, leading to feedback amplification, preservation of short-term metrics at the expense of structural integrity, or silent degradation masked by local correctness.

This work establishes prior art for the conceptual distinction between action-centric regulation and non-causal observation of viability and identity continuity. The contribution is architectural and conceptual rather than algorithmic. We delineate the problem space, characterize failure modes arising from causal entanglement between observation and control, and clarify why existing monitoring, safety, and optimization-based approaches are insufficient to address long-horizon structural degradation in adaptive systems.

1 Introduction

Adaptive systems operating in real-world environments must remain viable not only with respect to immediate task performance, but also with respect to their ability to sustain coherent organization over extended temporal horizons. Robots, autonomous agents, cyber-physical systems, and long-running software agents increasingly operate under conditions of partial observability, delayed feedback, non-stationary dynamics, and prolonged interaction, where failure does not arise from isolated incorrect actions but from cumulative structural effects unfolding over time.

Despite this reality, most existing adaptive architectures implicitly assume that preserving local correctness is sufficient for preserving long-term viability. As long as actions remain admissible, rewards remain stable, constraints are satisfied, and errors are bounded, the system is treated as healthy. This assumption is deeply embedded in control theory, reinforcement learning, safety frameworks, and optimization-based regulation, where evaluation is tightly coupled to action-level behavior and short-horizon performance metrics.

Empirical evidence across domains contradicts this assumption. Adaptive systems may remain locally correct while undergoing slow structural degradation that is invisible to action-centric metrics. Such systems often exhibit prolonged periods of apparently stable operation followed by abrupt failure, loss of adaptability, or collapse of organizational identity. In these cases, no single action is incorrect, no constraint is violated, and no immediate error signal indicates danger. The failure emerges because the system lacks a representation of its own long-horizon viability as a distinct object of observation.

This work argues that such failures arise from a conceptual and architectural limitation shared by existing approaches: the absence of a non-causal observational perspective on viability and identity continuity. In prevailing architectures, any evaluation of risk, degradation, or integrity is ultimately transformed into causal signals that influence behavior, learning, or optimization. Observation is inseparable from control, and long-horizon properties are collapsed into short-horizon decision processes.

As a consequence, adaptive systems are unable to observe their own long-term structural condition without simultaneously modifying their behavior. Viability becomes an implicit objective, identity becomes an optimization target, and observation becomes entangled with action. This entanglement prevents the system from forming a neutral, global view of its own organizational continuity and masks slow degradation processes until failure becomes unavoidable.

The present work isolates this limitation as a problem of architectural representation rather than of algorithmic deficiency. It identifies the need for an observational dimension in which viability and identity continuity can be assessed without becoming causally operative with respect to action selection, learning dynamics, or optimization objectives. The work does not propose a specific mechanism, controller, or implementation. Instead, it establishes prior art for the distinction between causal regulation and non-causal observation in the context of long-horizon adaptive viability.

This investigation is conducted within the broader context of the Petronus research project, which explores architectural principles for long-horizon sustainability, coherence, and identity preservation in adaptive systems. Within this framework, the present work serves to delineate the conceptual problem space by articulating why action-centric safety, control, and optimization mechanisms are insufficient to address silent structural degradation, and why a non-causal observational perspective is a necessary architectural axis for durable adaptive systems.

2 Viability and Identity as Long-Horizon Properties

In prevailing adaptive and control architectures, viability is most often treated as a form of constraint satisfaction or safety enforcement. A system is considered viable if its actions remain within admissible bounds, if constraints are not violated, and if optimization objectives continue to be met. When viability is formalized at all, it is typically encoded through action-level restrictions, penalty terms, or reward shaping mechanisms that operate on short temporal horizons.

Identity, by contrast, is rarely formalized explicitly. When it is addressed, it is commonly conflated with notions such as policy stability, parameter continuity, or similarity of observable

behavior across time. In such treatments, identity is reduced to surface-level persistence rather than understood as a structural property of organization.

Both approaches obscure a critical fact: viability and identity are inherently long-horizon properties. They are not properties of individual actions, single decisions, or instantaneous states. Instead, they emerge from cumulative interaction history, persistence of internal organization, and the system's ability to sustain coherent operation without requiring escalating corrective effort or increasingly brittle compensatory strategies.

A system may satisfy all local constraints and continue to produce acceptable outputs while its internal organization is progressively degraded. Structural flexibility may diminish, corrective interventions may become more frequent or more tightly coupled, and internal dependencies may accumulate in ways that reduce resilience. None of these processes is necessarily reflected in instantaneous reward signals, local error measures, or action-level safety checks.

Identity continuity, in this context, refers not to behavioral similarity but to the system's capacity to remain the same organized system over time. It captures whether the system continues to operate as a coherent whole, rather than fragmenting into a collection of narrowly compensated behaviors that only appear stable when viewed locally. A loss of identity may therefore occur even while task performance remains nominally correct.

Crucially, such degradation processes are typically silent. They unfold gradually and do not trigger alarms in architectures that rely exclusively on action-level metrics. Because no single action is erroneous and no immediate constraint is violated, the system appears healthy until a threshold is crossed and failure manifests abruptly. At that point, recovery is often impossible because the underlying organizational capacity has already been eroded.

This mismatch between what is measured and what ultimately fails reveals a conceptual gap in existing architectures. Viability and identity are treated as if they were short-horizon, action-level properties, when in fact they are global, temporally extended characteristics of system organization. Without an architectural perspective capable of representing these properties over long horizons, adaptive systems remain blind to the very conditions that determine their durability.

The remainder of this work builds on this observation by isolating the need for an architectural distinction between action-centric evaluation and long-horizon observation of viability and identity continuity, without prescribing how such observation is implemented or operationalized.

3 Causal Entanglement of Observation and Control

In existing adaptive architectures, degradation and risk are typically addressed by embedding evaluative mechanisms directly into the agent's control or learning loop. Signals derived from monitoring, diagnostics, or assessment are transformed into rewards, penalties, constraints, gradients, or auxiliary objectives that causally influence action selection and policy adaptation.

This design choice creates a structural entanglement between observation and control. Evaluation is not merely used to interpret system state, but becomes an active driver of behavior. While such coupling can be effective for enforcing short-horizon correctness, it introduces a set of systemic limitations that are architectural rather than algorithmic in nature.

First, observation becomes optimization. Once viability- or integrity-related information is made causally operative, the agent is incentivized to optimize with respect to the evaluative signal itself. The system begins to manage indicators rather than the underlying organizational conditions those indicators were intended to reflect. This phenomenon is not limited to explicit reward hacking; it arises whenever evaluative signals become part of the agent's objective landscape.

Second, feedback amplification emerges. Because evaluative signals influence behavior and behavior in turn influences evaluation, the system develops meta-dynamics in which it responds not only to the environment, but also to its own internal assessment processes. Over time, this can lead to oscillations, overcorrection, brittleness, or self-reinforcing loops that distort adaptation. These effects are especially pronounced when evaluation spans longer temporal horizons than the control cycle.

Third, silent failure modes persist. Even when evaluation is embedded causally, systems may learn to preserve the monitored metrics while underlying structural degradation continues unobserved. Compensatory behavior can mask loss of flexibility, increasing internal dependency, or erosion of organizational coherence. Because the evaluation mechanism is itself part of the causal loop, it may cease to function as an independent witness of degradation. Failure then appears sudden, not because it is abrupt, but because the architecture lacks a non-entangled perspective capable of detecting it.

These phenomena are not implementation errors, tuning failures, or deficiencies of specific algorithms. They arise from a fundamental architectural assumption: that long-horizon evaluation must influence behavior in order to be meaningful. When observation is made causal, it inevitably becomes entangled with control, optimization, and learning dynamics.

As a result, existing architectures lack a principled way to observe long-horizon viability and identity continuity without simultaneously altering the behavior being observed. This limitation motivates the need for an architectural perspective in which evaluation can exist without becoming an object of optimization, feedback, or causal influence.

4 The Need for Non-Entangled Long-Horizon Observation

The limitations described above point to a missing architectural capability in existing adaptive systems: the ability to observe and interpret long-horizon viability and identity continuity without becoming causally entangled with the behavior being observed.

Long-horizon properties such as viability degradation, structural exhaustion, or erosion of organizational identity unfold over extended temporal scales and depend on cumulative interaction history rather than instantaneous state or action quality. Meaningful observation of such properties therefore requires persistence, memory, and global interpretation. However, when such observation is embedded within the agent's decision-making loop, it inevitably becomes part of the causal machinery that generates behavior.

This creates an inherent conflict. If long-horizon evaluation is made actionable, it ceases to function as an independent observational perspective. If it is excluded entirely, the system lacks any representation of structural conditions that are not captured by local performance metrics.

Existing architectures resolve this conflict implicitly by prioritizing action-centric correctness, thereby sacrificing independent long-horizon observability.

A non-entangled observational perspective would allow viability and identity continuity to be assessed as properties of the system's ongoing organization, rather than as signals to be optimized. Such a perspective would not prescribe actions, impose constraints, or shape objectives. Instead, it would exist alongside behavior, maintaining a global view of structural evolution without exerting causal pressure on the agent's decision process.

The absence of such a perspective explains why many adaptive systems fail in ways that appear abrupt or inexplicable. Structural degradation accumulates outside the representational scope of action-level metrics, and because observation is causally entangled with control, no independent witness exists to register the decline. What is perceived as sudden failure is often the delayed manifestation of long-horizon processes that were never observed as such.

Accordingly, there exists a conceptual and architectural need for a form of long-horizon observation that is decoupled from action selection, optimization, and learning. Establishing the necessity of such non-entangled observation clarifies a previously under-articulated dimension of adaptive system design and delineates a problem space not addressed by existing control, safety, or learning frameworks.

This work situates that need within the broader context of the Petronus project, alongside related investigations into intervention cost, regime admissibility, and long-horizon viability. Together, these works articulate the boundaries of an architectural approach in which sustained operation is understood not as a byproduct of local correctness, but as a property requiring its own representational dimension.

5 The Missing Non-Causal Dimension

The failure modes described above point to a fundamental architectural absence in existing adaptive systems: the lack of a non-causal dimension for observing and interpreting long-horizon viability and identity continuity.

In prevailing architectures, observation is treated as either operationally inert or behaviorally active. On one extreme, systems may record logs, traces, or diagnostics that accumulate information without influencing operation. Such observation is passive and retrospective, offering no regulatory relevance during execution. On the other extreme, observation is embedded directly into control or learning mechanisms, where evaluative information is transformed into rewards, penalties, constraints, gradients, or decision rules that causally shape behavior.

There is no widely recognized architectural space between these extremes. Existing frameworks do not provide a principled way to observe long-horizon structural properties while keeping that observation non-causal with respect to action selection, learning dynamics, or optimization processes.

As a result, adaptive systems face an implicit trade-off. If long-horizon viability and identity are not observed, degradation proceeds unnoticed until failure. If they are observed through causal channels, the system's behavior becomes distorted by the act of observation itself. Evaluation collapses into control, and the observed quantities cease to reflect the system's unperturbed

structural condition.

This absence is not a limitation of specific algorithms or implementations. It is an architectural gap. Without a non-causal observational dimension, adaptive systems lack the capacity to maintain an external, long-horizon perspective on their own viability and identity without altering the very processes being observed.

6 Silent Degradation and Identity Loss

One of the most critical consequences of this architectural gap is the emergence of silent degradation. Adaptive systems may continue to operate correctly for extended periods while undergoing gradual deterioration of their internal organization. Actions remain admissible, rewards remain stable, and errors remain bounded, creating the appearance of healthy operation.

Because degradation unfolds over long temporal horizons, it is often invisible to action-level metrics and short-term performance indicators. No single decision or state transition signals failure. Instead, the system's capacity to sustain its organization erodes incrementally, without triggering alarms or corrective responses.

In this context, identity loss does not necessarily manifest as abrupt malfunction. Rather, it appears as a gradual transformation of the system into a more brittle or constrained form. The system may become increasingly dependent on compensation, tighter regulation, or reduced flexibility in order to maintain acceptable behavior. Over time, adaptability diminishes, and the system's ability to respond coherently to novel or unforeseen conditions is compromised.

Eventually, failure may occur suddenly, not because of a single catastrophic event, but because accumulated degradation has exhausted the system's capacity to absorb further disturbance. From the perspective of conventional monitoring, such failures appear unexpected, as no action-level metric directly reflected the underlying process.

These silent degradation modes cannot be reliably detected or prevented through action-centric safety mechanisms alone. As long as correctness is evaluated locally and causally tied to behavior, long-horizon structural decline remains outside the representational scope of the system. Addressing this class of failure therefore requires a fundamentally different perspective—one that separates the observation of long-term viability and identity from the mechanisms that generate behavior.

7 Scope and Non-Disclosure Boundary

This work deliberately refrains from proposing any specific mechanism, algorithm, architectural component, or system design for implementing non-causal observation of viability or identity continuity. No disclosure is made regarding how such observation might be realized, how long-horizon structural properties could be represented, stored, or evaluated, or how any form of assessment might be operationally integrated into an adaptive system.

In particular, this work does not describe how observation may be decoupled from action, how causality may be prevented in practice, or how any form of non-causal evaluation could be safely composed with control, learning, or decision-making processes. All questions of implementation,

embodiment, and system integration are intentionally left open.

The purpose of this work is strictly to establish prior art for the existence, relevance, and architectural significance of the problem itself. Specifically, it identifies that long-horizon viability and identity continuity cannot be adequately addressed within architectures where observation is causally entangled with action selection, learning, or optimization.

Any system that introduces a non-causal observational perspective on viability or identity—regardless of how such a perspective is implemented—operates within the conceptual and problem space delineated by this work. The contribution of this disclosure is therefore limited to problem identification and architectural framing, rather than solution specification.

Implications for Adaptive System Design

The analysis presented in this work has direct implications for the design and evaluation of adaptive systems intended for long-horizon operation. It suggests that durability, sustainability, and identity preservation cannot be guaranteed solely through incremental improvements in control accuracy, optimization techniques, reward shaping, or action-level safety constraints.

As long as viability-related evaluation is embedded causally within the same structures that generate behavior, adaptive systems remain vulnerable to distortion, feedback amplification, and silent degradation. Improvements in optimization may delay failure, but do not resolve the underlying architectural tension between observation and control.

The implications therefore extend beyond algorithm choice or performance tuning. They point to a need for architectural reconsideration: a separation between the processes that act within the environment and the processes that interpret long-horizon structural conditions of the system itself.

Without such separation, adaptive systems face a persistent dilemma. Either long-horizon viability and identity remain unobserved, allowing degradation to proceed unnoticed, or they are observed in a way that alters behavior and undermines the validity of the observation. Recognizing this dilemma is a prerequisite for any future architectural approach to long-horizon adaptive system design.

8 Conclusion

This work identifies a fundamental conceptual limitation inherent in existing adaptive system architectures: the absence of a non-causal perspective from which long-horizon viability and identity continuity can be observed without influencing behavior.

By isolating this limitation, the work establishes prior art for the architectural necessity of a non-causal observational dimension distinct from action selection, learning, and optimization. The contribution is explicitly not algorithmic or implementational. Instead, it clarifies a structural gap that cannot be resolved through improved control accuracy, richer rewards, or more sophisticated safety mechanisms alone.

The analysis demonstrates that action-centric monitoring, safety, and optimization frameworks are inherently insufficient to detect or prevent silent structural degradation and gradual identity loss in long-horizon operation. As long as observation remains causally entangled with behavior, adaptive systems remain vulnerable either to unobserved degradation or to behavioral

distortion driven by self-monitoring.

Recognizing the separation between causal control and non-causal long-horizon observation is therefore a prerequisite for the design and evaluation of adaptive systems intended to operate coherently and sustainably under real-world uncertainty.

MxBv, 2026
10.5281/zenodo.18185321