When Performance Persists but Identity Fails: A Conceptual Framework for Long-Horizon Adaptive Systems

Maksim Barziankou, Poznan 2026

Abstract

Identity Continuity as a Missing Dimension of Evaluation

Adaptive systems are increasingly deployed in environments characterised by uncertainty, non-stationarity, delayed feedback, and cumulative stress. In such settings, failure rarely manifests as abrupt breakdown or immediate loss of functionality. Instead, systems often continue to operate, respond, and optimise while undergoing gradual internal degradation that remains invisible to standard performance-based evaluation metrics.

This work introduces identity continuity as a foundational yet largely unmeasured property of adaptive systems. We argue that prevailing paradigms in reinforcement learning, control theory, and embodied intelligence implicitly assume preservation of internal identity throughout operation, and that this assumption becomes invalid in long-horizon and high-stress environments. Under such conditions, adaptive systems may undergo irreversible internal regime transitions while maintaining outwardly acceptable or even improved task-level performance.

We propose reframing robustness as a survival problem, defined by the duration over which a system preserves coherent internal organisation rather than by averaged performance indicators such as error, reward, or task success. Identity collapse is treated as a distinct failure mode, fundamentally different from transient degradation, instability, or optimisation inefficiency.

The contribution is intentionally conceptual and evaluative. No algorithms, mechanisms, thresholds, or implementation details are disclosed. Instead, the work establishes a principled foundation for identity-aware evaluation of adaptive systems and exposes structural blind spots in existing evaluation practices.

The framework described herein may be applied as a standalone evaluative layer or embedded within larger multi-layer architectures, including but not limited to PETRONUS. This publication is intended as prior art and as a conceptual reference point for future research on long-horizon robustness, safety, and identity-preserving adaptive systems.

1 Introduction: The Hidden Failure Mode of Adaptive Systems

Adaptive systems are traditionally evaluated through performance-oriented criteria such as error minimisation, reward accumulation, stability margins, or task completion rates. These metrics are well suited to controlled settings, short operational horizons, and environments in which failure manifests as an observable deviation from expected behaviour.

However, as adaptive systems are increasingly deployed in open-ended, non-stationary, and long-horizon environments, a qualitatively different class of failure emerges. In such settings,

systems may continue to act, optimise, and satisfy externally defined objectives while undergoing internal changes that are neither captured nor penalised by conventional evaluation metrics.

In practice, an adaptive system may remain responsive and outwardly competent while having silently transitioned into a different internal regime. Behavioural outputs may remain admissible, trajectories may remain bounded, and rewards may continue to accumulate. Yet the internal organisation responsible for prior behaviour may no longer exist. From an external perspective, the system appears intact. From an internal perspective, the system is no longer the same entity.

This discrepancy reveals a fundamental limitation of existing evaluation paradigms: they assess what a system does, but not what the system remains. Identity continuity—the preservation of coherent internal organisation over time—is assumed everywhere and measured nowhere.

This work addresses this gap by introducing identity continuity as an explicit evaluative dimension for adaptive systems. We argue that many phenomena commonly attributed to model drift, distribution shift, unexpected behaviour, or brittleness are more accurately understood as instances of identity discontinuity: irreversible internal regime transitions that occur without immediate collapse of externally measured performance.

By focusing on identity rather than outcome, this work exposes a hidden failure mode of adaptive systems—one that becomes increasingly dominant as operational horizons lengthen, environments destabilise, and cumulative stress replaces instantaneous perturbation as the primary driver of system degradation.

2 The Implicit Assumption of Identity Continuity

Most contemporary frameworks for adaptive systems rely on an implicit assumption: that the agent evaluated at the end of an operational episode is, in an internal and organisational sense, the same agent that began it. In other words, prevailing evaluation practices presuppose continuity of identity as a background condition, not as a property to be tested.

This assumption is rarely stated explicitly because, in short-horizon or well-posed settings, it is usually safe. When environments are stationary, feedback is timely, perturbations are limited, and the operational horizon is brief, internal organisation tends to remain sufficiently stable for performance metrics to be meaningful proxies of system quality.

Across major paradigms, the same presupposition appears in different forms.

In reinforcement learning, identity continuity is typically treated as a given within an episode. Episodes terminate based on external criteria such as task completion, time limits, or constraint violation. Unless an explicit reset or reinitialisation is invoked, the agent is evaluated as though its internal organisation remains continuous while it accumulates experience, adapts, and updates.

In classical control, bounded error, stability, or convergence are commonly interpreted as evidence that the controlled system remains intact in the relevant operational sense. While control theory addresses stability of trajectories and responses, it generally does not distinguish between recoverable deviation and internal regime changes that preserve bounded outputs while altering the organising structure that produces them.

In embodied and interactive agent architectures, continued interaction is often treated as a sufficient condition for continuity of the perception–action loop. As long as the loop remains operational and produces admissible behaviour, the underlying organisation is implicitly treated as preserved.

These assumptions can be valid under idealised conditions. They become fragile under prolonged exposure to noise, drift, delayed feedback, non-stationarity, or adversarial environments. In such settings, an adaptive system may preserve outward competence while its internal organisation degrades, fragments, or transitions into a qualitatively different regime.

The consequence is a blind spot: identity continuity is presupposed by most evaluation protocols, yet it is not directly measured. As operational horizons lengthen and cumulative stress dominates failure dynamics, this unmeasured variable increasingly determines whether performance metrics remain meaningful or become misleading.

3 Identity as an Operational Property

In this work, identity is not approached as a philosophical, psychological, or subjective notion. It is defined strictly in operational terms, as the persistence of an internal organisation that gives rise to characteristic patterns of behaviour over time.

Identity, in this sense, is not inferred from intention, awareness, or self-modeling. It is inferred from structural continuity. An adaptive system is said to preserve its identity insofar as the internal organisation responsible for generating its behaviour remains coherent and functionally continuous under perturbation.

This distinction is critical because external behaviour alone is insufficient to establish identity continuity. Two adaptive systems may produce indistinguishable outputs while relying on fundamentally different internal organisations. Likewise, a single system may continue to satisfy external criteria while having undergone an irreversible internal transition that alters how behaviour is generated, stabilised, or adapted.

Operational identity is therefore not reducible to task success, accuracy, reward accumulation, or error minimisation. These quantities describe outcomes. Identity describes the persistence of the internal structure that produces those outcomes. Once that structure changes irreversibly, behavioural continuity no longer implies identity continuity.

4 Coherence as the Carrier of Identity

Identity continuity manifests through coherence. Coherence refers to the internal alignment among the components of an adaptive system responsible for perception, interpretation, action generation, and temporal coordination.

Coherence is not a performance measure and is not optimised directly. It functions as an invariant of internal organisation: when coherence is preserved, the system remains internally consistent even as behaviour adapts. When coherence degrades, the internal organisation becomes strained, fragmented, or misaligned, regardless of whether external performance remains acceptable.

A key property of coherence is that its degradation typically precedes observable failure. Long before error increases, rewards diminish, or tasks fail, coherence may already be eroding through accumulated misalignment, excessive intervention, or regime instability. At regime transition boundaries, coherence may collapse abruptly, marking an irreversible change in internal organisation.

Importantly, coherence collapse does not require immediate loss of functionality. An adaptive system may continue to act, respond, and even perform well on externally defined objectives while its internal organisation has already transitioned into a different regime. In this sense, coherence serves as the carrier of identity continuity, revealing internal failure modes that remain invisible to outcome-based evaluation.

By treating coherence as an operational property rather than a performance signal, this work separates the question of *what* a system achieves from the question of *how* it continues to exist as the same system over time.

5 Regime Transitions and Irreversibility

Adaptive systems routinely experience transient degradation, including noise-induced deviation, temporary instability, or recoverable error. Such events do not threaten identity continuity.

However, there exist regime transitions that are qualitatively distinct. In these transitions, the internal organisation responsible for prior behaviour ceases to exist or is replaced by a fundamentally different structure. These transitions are irreversible in the operational sense: recovery of behaviour does not imply recovery of identity.

Distinguishing irreversible regime collapse from transient deviation is essential, yet remains unaddressed in existing evaluative frameworks.

6 Why Error and Reward Mask Identity Collapse

Error- and reward-based metrics characterise adaptive systems through aggregated outcomes. By construction, such metrics compress extended temporal behaviour into scalar summaries, emphasising average performance while suppressing information about internal evolution.

This aggregation has a critical side effect: rare but structurally decisive events are diluted, and gradual internal degradation is rendered statistically insignificant. A system may therefore appear stable, improving, or convergent while its internal organisation is progressively reconfigured through repeated intervention, adaptation, or regime switching.

In such cases, optimisation does not merely fail to prevent identity loss. It actively contributes to it. By reinforcing outcome-equivalent behaviours without regard to their internal cost, optimisation pressures may drive the system toward regimes that satisfy external criteria at the expense of internal coherence.

From the standpoint of identity continuity, these systems have failed, even when conventional metrics indicate success. Error and reward do not merely fail to detect identity collapse; they systematically conceal it.

7 Robustness as Survival, Not Performance

This work proposes reframing robustness as a survival problem rather than a performance problem.

Instead of asking how well a system performs according to predefined criteria, robustness is defined as the duration over which the system preserves coherent internal organisation under cumulative stress, uncertainty, and perturbation. The central quantity of interest is not correctness at a moment in time, but persistence of identity across time.

Under this framing, robustness reflects resistance to irreversible internal transition rather than resilience to transient disturbance. A system is robust insofar as it continues to exist as the same system, regardless of fluctuations in external performance.

This perspective is conceptually aligned with survival analysis in reliability engineering and biological systems, where longevity and failure modes are studied independently of momentary efficiency. However, it remains largely absent from the evaluation of adaptive systems, which predominantly emphasise averaged outcomes.

Importantly, the present work does not introduce a survival mechanism or prescribe how survival is to be achieved. Survival is used strictly as an evaluative lens, redefining what it means for an adaptive system to endure in long-horizon, high-stress environments.

8 Implications for Evaluation and Safety

The absence of mechanisms for detecting identity collapse has direct consequences for how adaptive systems are evaluated, deployed, and safeguarded.

Most safety frameworks are oriented toward preventing externally observable harm, constraint violations, or unsafe actions. They intervene when a system's outputs exceed predefined limits, violate formal constraints, or produce unacceptable outcomes. Such approaches implicitly assume that preserving correct external behaviour is sufficient to ensure system integrity.

Similarly, common evaluation practices emphasise convergence, stability, or bounded deviation. Early stopping criteria halt learning when performance plateaus or degrades, while monitoring systems track error magnitudes or reward trends. None of these approaches are designed to detect qualitative changes in internal organisation.

As a consequence, adaptive systems may undergo irreversible internal transitions while remaining externally compliant. Degraded, brittle, or misaligned internal regimes can persist unnoticed, accumulating latent risk that manifests only under rare or extreme conditions. From a safety perspective, this constitutes a silent failure mode.

Identity-aware evaluation reframes safety as a property of internal continuity rather than external correctness. Preserving coherent internal organisation becomes a prerequisite for reliable operation, not an optional refinement. In this sense, identity continuity is not a performance enhancement but a foundational safety requirement for long-horizon adaptive systems.

9 Relation to Existing Paradigms

The problem of identity continuity cuts across established paradigms without being explicitly addressed by any of them.

Reinforcement learning systems optimise expected reward over time and typically define episode boundaries through task completion, time limits, or externally imposed resets. Identity is assumed to persist implicitly between updates, and regime collapse is indistinguishable from recoverable degradation unless it directly impacts reward.

Classical control theory focuses on stability, convergence, and bounded error relative to reference trajectories. While these criteria ensure admissible behaviour within a specified operating envelope, they do not distinguish between transient deviation and irreversible internal reconfiguration, particularly when multiple internal realisations produce equivalent external responses.

Embodied agent architectures emphasise continuous interaction between perception and action, often grounded in geometric or sensorimotor consistency. However, sustained interaction is commonly taken as evidence of preserved identity, even when internal coordination degrades or reorganises in ways that compromise long-term viability.

Across these paradigms, identity continuity is assumed rather than modelled, and regime transitions are treated as secondary effects rather than primary objects of analysis. The framework advanced in this work does not replace existing approaches, but exposes a shared blind spot: the lack of conceptual tools for recognising when an adaptive system ceases to remain itself.

10 Architectural Scope and Applicability

The framework described herein is intentionally defined at an architectural and evaluative level. It may be realised as a standalone evaluative layer operating alongside existing adaptive systems, or it may be embedded within broader multi-layer architectures that address additional concerns such as control, learning, or coordination.

One example of such a broader architecture is PETRONUS, in which identity-aware evaluation constitutes a distinct and separable layer concerned with long-horizon viability and internal continuity. The present work does not disclose how such integration is implemented, nor does it prescribe any specific interfaces or mechanisms by which the framework interacts with other architectural components.

The framework is designed to be architecture-agnostic. It does not depend on a particular learning paradigm, control law, embodiment, or representational scheme. In particular, it does not require explicit environmental models, geometric reconstructions, or task-specific state representations. Its applicability extends across heterogeneous environments and system designs, provided that some notion of internal organisation and temporal evolution is present.

11 What This Work Does Not Provide

This work deliberately refrains from specifying concrete mechanisms. It does not introduce algorithms, thresholds, detection rules, or computational procedures. No update equations,

parameter schedules, sensor dependencies, or optimisation strategies are disclosed.

The framework does not define how coherence is measured, how identity discontinuity is detected, or how evaluative signals are operationalised. These aspects are left unspecified by design. The intent is to establish a conceptual and evaluative foundation that remains independent of particular implementations and resistant to premature optimisation or overfitting to specific system classes.

By withholding implementation details, this work avoids constraining future realisations and preserves flexibility for diverse architectural interpretations. The contribution lies not in a prescribed solution, but in the identification of a neglected problem space and the articulation of a principled perspective from which that space may be addressed.

12 Conclusion

Adaptive systems do not fail solely by making mistakes. They fail by ceasing to preserve their internal identity.

This work has argued that identity continuity constitutes a fundamental yet largely unexamined dimension of adaptive behaviour. Prevailing evaluation paradigms implicitly assume that an adaptive system remains internally the same entity throughout its operation, even as it learns, adapts, or optimises. In long-horizon, uncertain, and high-stress environments, this assumption breaks down. Systems may continue to operate, respond, and even perform well while having undergone irreversible internal regime transitions that conventional metrics neither detect nor acknowledge.

By reframing robustness as a survival property—the duration over which coherent internal organisation is preserved—this work shifts the focus of evaluation away from instantaneous correctness and toward long-horizon viability. Identity discontinuity is treated not as a variant of task failure or instability, but as a qualitatively distinct mode of collapse that requires its own conceptual and evaluative treatment.

The framework introduced herein does not prescribe mechanisms, algorithms, or implementation details. Instead, it delineates a problem space and establishes a principled lens through which adaptive systems may be analysed, compared, and ultimately designed. In doing so, it exposes a class of silent failures that remain invisible to performance-based evaluation and highlights the need for identity-aware perspectives in safety, reliability, and long-term autonomy.

Future advances in adaptive systems will increasingly depend not on how effectively systems optimise outcomes, but on how reliably they preserve coherent internal organisation over time. Robustness, in this sense, is not a measure of success, but of continued existence.

A system is not robust because it performs well. It is robust because it remains itself.

MxBv, Poznań 2026.
10.5281/zenodo.18165785