Identity Continuity and Regime Collapse in Adaptive Systems

A Survival-Based Perspective Beyond Error and Reward

Barziankou Maksim (MxBv)
Poznań 2026

Abstract

Adaptive systems are traditionally evaluated using performance-oriented metrics such as error minimisation, cumulative reward, or task success rate. These metrics implicitly assume that the internal identity of an agent remains continuous throughout its operation. In short-horizon or well-controlled settings, this assumption is often acceptable. In long-horizon, uncertain, or high-stress environments, however, it becomes fundamentally invalid.

In real-world operation, adaptive systems may undergo gradual or abrupt internal regime transitions in which the organisational structure responsible for prior behaviour degrades or irreversibly changes. Crucially, such transitions may occur without immediate task failure or significant degradation in externally observed performance. As a result, systems can appear functional, stable, or even optimised, while having silently lost the internal coherence that previously defined their behaviour.

This work introduces the notion of identity continuity as a foundational property of adaptive systems that is distinct from performance, accuracy, or reward. Identity continuity is treated as the persistence of a coherent internal organisation over time and under perturbation. Loss of identity continuity corresponds to an irreversible transition into a different internal regime, rather than a recoverable error or transient instability.

Building on this perspective, robustness is reframed not as sustained performance quality, but as a survival problem: the ability of a system to preserve coherent internal organisation over extended operation in the presence of uncertainty, noise, drift, and rare destabilising events. Robustness is thus naturally characterised by the duration of coherent existence prior to identity collapse, rather than by averaged or cumulative performance metrics.

The paper develops a conceptual and evaluative framework for identity-aware analysis of adaptive systems. It highlights the limitations of existing reinforcement learning, control-theoretic, and embodied-agent paradigms in detecting identity discontinuity and regime collapse. The contribution is intentionally architectural and conceptual in nature and does not propose specific algorithms, thresholds, or implementation mechanisms. Instead, it establishes a principled foundation for future work on identity-preserving adaptive systems and long-horizon robustness evaluation.

1 Introduction

1. Introduction / Problem Statement

Adaptive systems are increasingly deployed in environments characterised by uncertainty, non-stationarity, delayed feedback, and cumulative stress. In such settings, failure rarely manifests as an 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 evaluation metrics.

In practice, an adaptive system may continue to satisfy externally defined objectives while having silently transitioned into a qualitatively different internal state. From an external perspective, behaviour appears stable, competent, and goal-directed. From an internal perspective, however, the organisational structure that previously generated this behaviour may no longer exist. The system is operational, but it is no longer the same system.

This discrepancy exposes a fundamental blind spot in prevailing evaluation paradigms. Performance-based metrics implicitly conflate behavioural output with internal continuity, assuming that successful action implies preservation of the underlying agent. In long-horizon and high-stress environments, this assumption does not hold. Behavioural adequacy can persist long after the internal coherence that once supported it has collapsed or been replaced.

Such silent regime transitions are especially problematic in adaptive systems that learn, self-adjust, or accumulate structural changes over time. A system may compensate locally for degradation, mask instability through optimisation, or stabilise behaviour in a degraded regime that differs fundamentally from the original. Conventional metrics smooth over these transitions through averaging, reward accumulation, or convergence criteria, rendering identity loss undetectable.

This work addresses a foundational gap in current approaches to adaptive systems: the absence of conceptual and evaluative tools for detecting when an agent ceases to preserve its internal identity. The focus is not on task failure, safety violations, or performance degradation, but on a deeper phenomenon—the irreversible loss of internal organisation that defines "being the same system over time."

By separating identity continuity from performance outcomes, the paper argues for a reframing of robustness as an existential property rather than an optimisation result. The goal is not to propose specific detection mechanisms or implementations, but to establish a principled conceptual foundation for identity-aware analysis of adaptive systems operating under uncertainty.

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 internally identical to the agent that began it. This assumption is rarely stated explicitly, yet it underlies the majority of evaluation, optimisation, and safety criteria used in practice.

In reinforcement learning, identity continuity is assumed across time steps and episodes unless an explicit reset is applied. Learning dynamics, convergence analysis, and performance guarantees all presuppose that the same underlying agent persists throughout training and deployment. Classical control theory similarly assumes that bounded error, stability, or convergence imply preservation of the controlled system's internal structure. As long as trajectories remain bounded, the system is treated as fundamentally unchanged.

Embodied agent architectures extend this assumption to the perception–action loop. As long as sensing, actuation, and feedback remain feasible, continuity of the agent is presumed. Breakdowns are typically framed in terms of sensor failure, actuator limits, or environmental mismatch, not loss of internal identity.

These assumptions hold in idealised, short-horizon, or low-stress settings. They begin to fail under prolonged exposure to noise, drift, delayed feedback, partial observability, or adversarial conditions. In such environments, adaptive systems may undergo gradual or abrupt internal reorganisation that is not captured by performance metrics, stability criteria, or reward signals.

Crucially, identity continuity is assumed everywhere yet measured nowhere. Existing frameworks lack conceptual variables, observables, or evaluation protocols that would allow detection of when an agent ceases to be the same system in an internal, organisational sense. As a result, regime collapse, structural substitution, or irreversible degradation may occur silently, masked by acceptable outward behaviour.

3 Identity as an Operational Property

In this work, identity is not treated as a philosophical, cognitive, or subjective construct. It is neither equated with consciousness nor tied to semantic notions of agency or intention. Instead, identity is defined operationally as the persistence of an internal organisation that gives rise to consistent behavioural characteristics over time.

Operational identity refers to the continued existence of a coherent internal structure capable of generating behaviour in a stable and self-consistent manner under perturbation. This definition deliberately avoids reliance on specific representations, parameters, or models. Identity is not a stored object but a maintained organisation.

Two adaptive systems may exhibit indistinguishable external behaviour while differing fundamentally in their internal organisation. Conversely, a single system may preserve outward behavioural adequacy while undergoing an irreversible internal transition that replaces the organisational structure responsible for that behaviour. In both cases, external performance alone is insufficient to determine whether identity has been preserved.

Identity, in this sense, is not reducible to task success, accuracy, reward accumulation, or error minimisation. A system may optimise successfully while losing the very internal coherence that once made such optimisation possible. Identity is therefore tied not to outcomes, but to the structure and alignment of internal processes over time.

By treating identity as an operational property, this work shifts the focus from what a system achieves to how it continues to exist as the same system while achieving it. This distinction becomes critical in long-horizon and high-stress environments, where survival of internal organisation cannot be inferred from short-term behavioural success.

4 Coherence as the Carrier of Identity

Identity continuity manifests through coherence: the internal alignment among perception, interpretation, action generation, and temporal coordination. Coherence is the property that allows an adaptive system to behave as a unified whole rather than as a collection of loosely coupled processes.

Coherence is not a performance metric and cannot be inferred from external success alone. It does not measure how well a task is solved, how accurate a prediction is, or how large a reward is accumulated. Instead, coherence reflects whether internal processes remain mutually compatible, synchronised, and structurally aligned over time.

Empirically, coherence typically degrades prior to observable failure. Long before error explodes or performance collapses, internal alignment weakens: timing relationships drift, internal signals lose mutual consistency, and corrective actions become increasingly fragmented. At regime transition boundaries, coherence collapses abruptly, marking a qualitative change in internal organisation.

Crucially, coherence collapse may occur without immediate task failure. A system may remain accurate, responsive, or functionally adequate while internally fragmenting. External behaviour may appear stable even as the internal structure that once generated that behaviour ceases to exist.

A familiar analogy can be observed in human behaviour. A person may continue to perform routine actions—walking, speaking, executing learned procedures—while experiencing loss of internal integration, disorientation, or dissociation. Behaviour persists, yet the underlying coherence that once unified perception, intention, and action has been compromised.

In adaptive systems, coherence serves as the operational carrier of identity. When coherence is preserved, identity persists. When coherence collapses irreversibly, identity is lost regardless of outward performance.

5 Regime Transitions and Irreversibility

Adaptive systems routinely experience transient degradation. Noise-induced deviation, temporary instability, delayed feedback, partial observability, or recoverable error are inherent features of real-world operation. Such events may degrade performance temporarily but do not threaten identity continuity.

In these cases, the internal organisation responsible for behaviour remains intact. Once perturbations subside or compensatory mechanisms engage, coherence is restored and the system continues as the same entity.

However, there exist regime transitions that are qualitatively different. In these transitions, the internal organisation responsible for prior behaviour ceases to exist or is replaced by a fundamentally different structure. The system may continue to operate, but it does so as a different internal entity.

These transitions are irreversible in the operational sense. Recovery of outward behaviour does not imply recovery of identity. A system may regain accuracy, stability, or task success while the internal structure that once supported such behaviour has been permanently altered or destroyed.

Irreversibility, in this context, does not imply physical damage or explicit failure. It denotes the loss of the internal organisational configuration that defined the prior identity. Once this configuration is lost, subsequent behaviour—regardless of similarity—belongs to a different internal regime.

Distinguishing irreversible regime collapse from transient deviation is essential for understanding robustness in adaptive systems. Yet existing frameworks lack concepts, observables, or evaluative criteria capable of making this distinction. As a result, irreversible identity loss is routinely misclassified as recoverable error, adaptation, or noise.

6 Why Error and Reward Mask Identity Collapse

Error-based and reward-based metrics are inherently averaging. They compress rich temporal dynamics into scalar summaries that prioritise aggregate performance over structural integrity. As a result, they smooth over rare but catastrophic events and obscure gradual internal degradation.

In long-horizon operation, identity collapse rarely manifests as an immediate spike in error or a sudden drop in reward. Instead, it unfolds as a slow erosion of internal organisation: compensatory mechanisms overactivate, internal signals decouple, and behavioural consistency is maintained through increasingly fragile means. Averaging metrics interpret this process as stability or adaptation.

A particularly dangerous scenario arises when optimisation actively accelerates identity loss. An adaptive system may optimise itself into a different internal regime that is better aligned with short-term reward or error minimisation, while abandoning the organisational structure that previously defined its identity. From the perspective of conventional metrics, performance improves. From the perspective of identity continuity, the original system has ceased to exist.

In such cases, optimisation does not prevent failure; it conceals it. Error and reward become instruments that legitimise regime substitution rather than indicators of robustness. Identity collapse is not detected because the metrics were never designed to observe it.

From the standpoint of identity continuity, these systems have failed—even when conventional evaluation frameworks declare success. This mismatch exposes a fundamental limitation of performance-centric assessment in adaptive systems.

7 Robustness as Survival, Not Performance

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

Instead of asking how accurately, efficiently, or optimally a system performs on average, robustness is understood as the duration for which the system preserves coherent internal organisation under cumulative stress. Robustness becomes a measure of resistance to irreversible identity collapse, not a measure of task proficiency.

Under this framing, time becomes the primary axis of evaluation. An adaptive system is considered robust if it can maintain identity continuity while exposed to noise, drift, delayed feedback, partial observability, and structural perturbations. Failure is defined not by poor performance, but by loss of coherent existence as the same system.

This perspective aligns naturally with survival analysis as used in reliability engineering, systems biology, and medicine, where interest lies not in average function but in time-to-failure under stress. Yet, despite its relevance, this framing remains largely absent from adaptive systems research, which continues to prioritise performance aggregates over existential persistence.

Importantly, this work introduces no specific survival mechanisms, recovery strategies, or control laws. Survival is defined purely as an evaluative construct. The contribution lies in establishing a conceptual lens through which robustness can be assessed independently of reward, error, or task success.

By separating evaluation from optimisation, this reframing exposes failure modes that remain invisible under traditional metrics and provides a foundation for identity-aware analysis of adaptive systems operating in real-world conditions.

8 Implications for Evaluation and Safety

The inability to detect identity collapse has profound implications for both evaluation and safety of adaptive systems.

Contemporary safety mechanisms are primarily oriented toward preventing externally observable harm. They enforce constraints on outputs, actions, or boundary conditions, while remaining agnostic to the internal organisational state of the system. As long as outward behaviour remains within acceptable limits, internal degradation is not treated as a safety concern.

Similarly, evaluation protocols rely on criteria such as convergence, bounded error, or sustained reward accumulation. Early stopping conditions optimise numerical stability rather than structural integrity. Monitoring systems track threshold violations in performance variables, not transitions between internal regimes.

As a consequence, adaptive systems may silently transition into degraded, brittle, or fundamentally unsafe internal organisations while remaining outwardly compliant. Such systems may satisfy formal safety constraints while having lost the internal coherence required for reliable operation under stress.

From this perspective, identity-aware evaluation is not a performance enhancement or an optional diagnostic tool. It represents a foundational safety requirement. Without the ability to detect loss of identity continuity, no adaptive system operating in open-ended environments can be meaningfully certified as safe over long horizons.

9 Relation to Existing Paradigms

Reinforcement learning frameworks optimise reward trajectories over episodes but lack explicit representation of identity boundaries. Episodes are treated as optimisation units rather than existential intervals, and regime collapse is indistinguishable from suboptimal exploration or policy drift.

Classical control theory provides rigorous tools for stability analysis and error bounding. However, stability in the control-theoretic sense does not imply preservation of internal organisation. A system may remain stable while having irreversibly transitioned into a different operational regime.

Embodied agent architectures emphasise interaction, embodiment, and sensorimotor coupling. Yet, continued interaction is typically equated with preserved identity. As long as perception–action loops remain closed, internal regime transitions are rarely considered.

Across paradigms, identity continuity is implicitly assumed rather than explicitly modelled. While each framework addresses aspects of robustness, none provides conceptual or evaluative tools for detecting when an adaptive system ceases to exist as the same system in an internal sense.

10 Toward Identity-Aware Evaluation Without Mechanisms

This work deliberately refrains from proposing specific mechanisms, algorithms, or architectural implementations for detecting identity discontinuity.

This choice is not a limitation but a methodological boundary.

The central claim of this paper is not that a particular solution exists, but that a fundamental evaluative dimension is missing from current approaches to adaptive systems. Before mechanisms can be meaningfully designed, the underlying property must be recognised, defined, and legitimised as an object of evaluation.

Identity continuity, as introduced here, is not a control variable, an optimisation target, or a behavioural objective. It is a precondition for meaningful operation over long horizons. Any mechanism introduced prematurely risks conflating identity preservation with performance optimisation, thereby reproducing the very failure modes this work seeks to expose.

By remaining mechanism-agnostic, this paper establishes identity continuity as an independent evaluative concept. It separates the question of whether an adaptive system remains itself from the question of how such preservation might be enforced, detected, or encouraged.

This separation is essential. Evaluation must precede intervention. Observation must precede control.

The absence of mechanisms in this work is therefore intentional and protective. It ensures that identity-aware evaluation cannot be reduced to a tuning parameter, reward proxy, or auxiliary loss. It also prevents premature optimisation pressure from distorting the boundary between recoverable degradation and irreversible regime collapse.

In this sense, the contribution of this paper is foundational rather than prescriptive. It provides a conceptual scaffold upon which future work may build mechanisms, operators, or architectures, while preserving a clear distinction between evaluation and optimisation.

The framework presented here is compatible with multiple paradigms, architectures, and implementation strategies. However, no particular approach is endorsed. The work establishes what must be measured before asking how it might be measured.

By formalising identity continuity as an evaluative primitive, this paper opens a new axis of analysis for adaptive systems—one that is orthogonal to performance, reward, accuracy, and task success, yet essential for long-horizon robustness and safety.

11 Open Questions and Research Directions

This work deliberately defines a problem space rather than a solution space.

By introducing identity continuity as an evaluative dimension, it exposes a set of questions that have remained structurally invisible within existing paradigms of adaptive systems. These questions are not incremental refinements of current methods, but foundational challenges that require new forms of thinking.

One open question concerns observability: how internal coherence can be inferred, approximated, or characterised without introducing optimisation pressure that distorts the very phenomenon under observation. Any attempt to measure identity risks altering it, and the boundary between observation and intervention remains unresolved.

A second question concerns irreversibility: how to distinguish transient degradation, noise-induced instability, or recoverable deviation from genuine regime collapse. This distinction cannot rely solely on magnitude, duration, or thresholding, and demands principled criteria grounded in internal organisation rather than external performance.

A third question concerns evaluation practice: how identity-aware evaluation can coexist with existing benchmarks, metrics, and validation protocols without being subsumed by them. Long-horizon identity loss may remain invisible in short-horizon tasks, raising questions about experimental design, stress testing, and comparative assessment.

These questions do not point toward a single method or architecture. Instead, they define a new research direction—one concerned not with how adaptive systems optimise, but with how they remain themselves over time.

12 Conclusion

Adaptive systems do not fail solely by making mistakes.

They fail by ceasing to preserve the internal organisation that once made their behaviour meaningful, stable, and interpretable. Such failure may occur silently, without dramatic error spikes, safety violations, or loss of task performance. From the outside, the system appears to function. Internally, it has become something else.

This work identifies identity continuity as a fundamental yet unmeasured property of adaptive systems. It argues that prevailing evaluation frameworks implicitly assume identity preservation while providing no means to verify it. As a result, systems may be declared robust precisely at the moment when robustness has already been lost.

By reframing robustness as survival of internal organisation rather than persistence of performance, this paper introduces a new axis of analysis that cuts across reinforcement learning, control theory, embodied intelligence, and autonomous systems. This axis is orthogonal to reward, accuracy, and task success, yet essential for long-horizon operation under uncertainty.

The contribution of this work is not a mechanism, an algorithm, or a design prescription. It is a conceptual realignment. It establishes identity continuity as an evaluative primitive and regime collapse as a distinct mode of failure—one that cannot be reduced to error, instability, or poor optimisation.

In doing so, the paper lays a foundation rather than a solution. It defines what must be preserved before asking how preservation might be achieved. Any future architecture that claims robustness without addressing identity continuity does so incompletely.

If adaptive systems are to operate safely, meaningfully, and reliably in the real world, they must be evaluated not only by what they achieve, but by whether they remain themselves while achieving it.

Author Statement

This work is intended as a conceptual and theoretical prior-art contribution. It deliberately avoids disclosure of specific mechanisms, architectures, or implementations.

MxBv, 2026

10.5281/zenodo.18153949