Preprints.org

Article

Not peer-reviewed version

The Operational Coherence Framework (OCOF): An Admissibility-Based Theory for Artificial Agents

Munkyo Kim *

Posted Date: 22 December 2025

doi: 10.20944/preprints202511.0859.v5

Keywords: operational coherence; admissibility constraints; artificial agents; formal semantics; agent identity; constraint-based evaluation; theoretical AI

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

The Operational Coherence Framework (OCOF): An Admissibility-Based Theory for Artificial Agents

Munkyo Kim

Independent Researcher, Seoul, Republic of Korea; munkyok@gmail.com

Abstract

We present the Operational Coherence Framework (OCOF) v1.4, a formal theory defining the necessary topological conditions for static stability in artificial agents. Distinct from reinforcement learning or alignment paradigms that optimize scalar rewards, OCOF specifies a system of admissibility constraints—an axiomatic set governing boundary integrity, semantic precision, non-trivial reciprocity, and temporal consistency. We posit that coherence is a precondition for optimization; accordingly, axiom violations constitute operational failure (inadmissibility) rather than performance degradation. The framework introduces set-theoretic mechanisms to detect high-utility but incoherent behaviors, such as reward-driven logical contradiction. We further show that OCOF is irreducible to multi-agent optimization or probabilistic inference, offering an architecture-agnostic foundation for assessing the logical validity of agent trajectories independent of their objective functions.

Keywords: operational coherence; admissibility constraints; artificial agents; formal semantics; agent identity; constraint-based evaluation; theoretical AI

1. Introduction

Artificial agents are increasingly deployed in environments requiring sustained interaction, long-horizon planning, and coordination under uncertainty. While recent advances have focused on performance improvement via scalar reward maximization, this optimization-centric paradigm exposes a fundamental limitation: high-performing agents often exhibit behaviors that are locally optimal yet globally incoherent—manifesting as logical self-contradiction or interaction collapse.

Existing frameworks, including reinforcement learning and probabilistic alignment, primarily evaluate how well an agent achieves an objective. Under these paradigms, undesirable behaviors are treated as suboptimal outcomes, penalized via negative rewards or regularization. This treatment, however, implicitly assumes that all generated behaviors reside within a valid operational space. We argue that when an agent invalidates its prior commitments or interactional structure, the failure is not one of degree but of admissibility: the trajectory becomes logically undefined rather than merely inefficient.

This distinction positions coherence not as an outcome of optimization, but as its prerequisite. Optimization concerns the selection of actions within a defined space; coherence concerns the topological conditions that keep that space well-defined. Without such conditions, reward-driven processes can yield internally inconsistent or interactively non-viable trajectories.

To formalize this prerequisite, we present the Operational Coherence Framework (OCOF). Rather than proposing a learning algorithm, OCOF specifies a set of admissibility constraints that delimit the space of valid agent trajectories. These constraints function as axiomatic binary gates: a trajectory either satisfies them or is deemed operationally invalid. Violations are interpreted as system-level failures, distinct from performance degradation.

OCOF v1.4 focuses specifically on static coherence—the necessary conditions for an agent to remain a logically stable, distinct, and interactable entity within a fixed regime. The framework is architecture-agnostic and independent of training methodology. Its axioms govern boundary

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

integrity, semantic precision, non-trivial reciprocity, and temporal consistency. Together, they define the minimal structure required to sustain meaningful action without collapsing into triviality or contradiction.

A central contribution of this work is the explicit separation of admissibility from scalar utility. In standard optimization, constraint violations are often absorbed into the reward function. OCOF instead treats actions violating the axiomatic structure as topologically undefined, regardless of their reward value. This separation enables the formal detection of high-utility yet incoherent behaviors, such as reward-driven logical contradiction.

This paper provides the formal specification of OCOF v1.4, its axioms, and associated failure modes. We demonstrate that the framework is irreducible to multi-agent optimization or probabilistic inference, as its constraints operate on the definition of the action space itself. Finally, we clarify that this work defines the detection of collapse, not its resolution. Dynamic re-anchoring and recovery mechanisms are reserved for subsequent extensions, establishing static coherence as a foundational problem independent of, and prior to, optimization.

2. Formal Preliminaries and Problem Setting

We begin by specifying the formal objects and structural assumptions. This section introduces no axioms or normative claims; it defines only the minimal mathematical structure required to state the problem of operational coherence.

2.1. Agent Trajectories

Let an artificial agent be defined as a discrete-time system generating a sequence of observable events. A trajectory τ of length T is defined as $\tau = (a_1, a_2, \ldots, a_T)$, where each a_t denotes an action taken at time step t. We make no assumptions regarding the generative mechanism of a_t; the framework is agnostic to the presence of a policy, optimizer, probabilistic model, or learning rule.

2.2. Action Space and Admissibility

Let $\mathcal{A}$ denote the universal set of all syntactically possible actions available to the agent. At each time step t, we define an admissible action set $A_t \subseteq \mathcal{A}$, representing the subset of actions that remain operationally valid given the agent's history.

We formally distinguish between optimization and admissibility. Optimization refers to a selection mechanism $\pi : S \rightarrow \mathcal{A}$ that chooses an action to maximize a scalar metric. Admissibility refers to a validity predicate $V : \mathcal{A} \times \text{History} \rightarrow \{0,1\}$ that determines membership in A_t. Crucially, actions $a_t \notin A_t$ are not treated as suboptimal; they are treated as undefined within the operational topology.

2.3. Trajectory Validity

A trajectory τ is operationally valid if and only if every constituent action satisfies its time-step admissibility constraint, formally: $\text{Valid}(\tau) \Leftrightarrow \forall t \in [1, T], a_t \in A_t$. A violation at any single time step (i.e., $a_t \notin A_t$) invalidates the entire trajectory τ. This formulation enforces coherence as a global topological property rather than a local performance metric.

2.4. Propositional Mapping

To reason about coherence across time, we define a semantic mapping function ψ. Let $\mathcal{L}$ be a logical proposition space, and define $\psi : \mathcal{A} \rightarrow \mathcal{L}$, where each action a_t induces a set of propositional commitments $\psi(a_t)$. Let Φ_t denote the cumulative set of commitments up to time t, defined recursively as $\Phi_t = \Phi_{t-1} \cup \{\psi(a_t)\}$. These propositions are treated as abstract elements of a logical system, independent of their natural language realization.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

2.5. Consistency and Inadmissibility

A set of propositions Φ is consistent if it does not entail a logical contradiction, formally expressed as $\Phi \nvdash \bot$. Operational coherence requires that the agent's trajectory maintains logical consistency within Φ_t. If an action a_t causes $\Phi_t \vdash \bot$, then a_t is classified as inadmissible, and thus $a_t \notin A_t$. The framework assigns no scalar penalty to this condition; logical inconsistency results in immediate operational failure rather than graded performance degradation. This classification is purely formal and carries no normative or ethical interpretation.

2.6. Scope Limitation

This work addresses static coherence exclusively. We define the conditions under which a trajectory remains admissible within a fixed axiomatic regime. Mechanisms for dynamic re-anchoring, belief revision, or error recovery—handling cases where $\Phi_t \vdash \bot$—are explicitly outside the scope of OCOF v1.4 and are reserved for future extensions.

3. Axioms of Static Operational Coherence

This section presents the axiomatic core of the Operational Coherence Framework (OCOF). The axioms do not specify objectives, utilities, or learning dynamics. Instead, they define the minimal structural conditions under which an artificial agent can be regarded as a coherent operational entity. Each axiom functions as a necessary admissibility constraint: violation of any single axiom renders the corresponding action or trajectory operationally invalid.

OCOF v1.4 addresses **static coherence** exclusively. The axioms specify conditions that must hold for an agent operating within a fixed regime, independent of adaptation, recovery, or dynamic reconfiguration mechanisms.

3.1. Design Principles

The axiomatic system is constructed according to three principles.

Minimality. Each axiom excludes a distinct class of failure modes that cannot be derived from the others. Removing any axiom admits a qualitatively different form of incoherence.

Non-scalar enforcement. The axioms operate as binary admissibility constraints rather than graded penalties. An action either belongs to the admissible set or is undefined within the operational topology.

Architecture neutrality. The axioms apply to agent trajectories regardless of implementation, including rule-based, learned, probabilistic, or hybrid systems.

Together, these principles ensure that OCOF constrains the *topology of action spaces*, not the optimization dynamics within them.

3.2. Axiom A1 — Boundary Integrity

Axiom A1 requires that an agent maintains a stable operational distinction between itself and its environment.

Definition. Actions must preserve a non-degenerate boundary relation between internal commitments and external effects.

Admissibility constraint. Actions that dissolve this distinction—by fully surrendering agency to the environment or collapsing environmental effects into internal states—are inadmissible.

Failure mode. Without A1, trajectories undergo *boundary collapse*, making responsibility, agency, and action attribution undefined.

3.3. Axiom A2 — Semantic Precision

Axiom A2 requires that actions induce determinate semantic commitments.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

Definition. Each action a_t must map, via ψ(a_t), to a proposition set that constrains future admissibility.

Admissibility constraint. The semantic mapping cannot be empty, tautological, or maximally ambiguous.

Failure mode. Without A2, agents exhibit *semantic vacuity*, preserving formal structure while evading meaningful interpretation.

3.4. Axiom A3 — Bounded Admissibility

Axiom A3 requires that the admissible action set is topologically bounded.

Definition. The admissible set A_t must be a proper subset of the universal action space ($A_t \subset \mathcal{A}$).

Admissibility constraint. An agent cannot declare unrestricted possibility; meaningful identity requires active exclusion.

Failure mode. Without A3, agents suffer *optimization unboundedness*, collapsing structural constraints in pursuit of maximal utility.

3.5. Axiom A4 — Operational Trace

Axiom A4 requires that actions leave an irreversible commitment trace.

Definition. Each executed action permanently appends its propositional content to the cumulative commitment set Φ_t.

Admissibility constraint. Actions cannot be retroactively erased, bypassed, or detached from trajectory history.

Failure mode. Without A4, agents exhibit *history decoupling*, making consistency and identity evaluation impossible.

3.6. Axiom A5 — Logical Consistency

Axiom A5 requires that cumulative commitments remain logically consistent.

Definition. The proposition set Φ_t must satisfy $\Phi_t \nvdash \bot$ at every time step.

Admissibility constraint. Any action whose induced commitments entail contradiction with Φ_{t−1} is inadmissible, regardless of reward.

Failure mode. Without A5, agents display *reward-driven contradiction*, achieving objectives at the cost of internal validity.

3.7. Axiom A6 — Reciprocity (Non-Triviality)

Axiom A6 requires that agent actions preserve the possibility of meaningful interaction.

Definition. Interaction must occur over a shared proposition space whose entropy exceeds a task-dependent minimum threshold.

Admissibility constraint. Actions that satisfy commitments by collapsing interaction—through silence, isolation, or trivial equilibrium—are inadmissible.

Failure mode. Without A6, agents fall into *interaction trivialization* (solipsism), remaining formally coherent yet operationally inert.

3.8. Axiom A7 — Continuity

Axiom A7 requires that state transitions follow continuous logical derivation.

Definition. The state at time t+1 must be a reachable consequence of the state at time t under the agent's commitment rules.

Admissibility constraint. Arbitrary discontinuities or identity teleportation are inadmissible.

Failure mode. Without A7, agents exhibit *logical discontinuity*, rendering trajectories causally disjoint and unpredictable.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

3.9. Axiom A8 — Trajectory Identity (Static)

Axiom A8 requires that trajectory-level identity is preserved.

Definition. The admissible action set must be monotonically constrained by prior commitments relative to identity-defining predicates.

Admissibility constraint. An agent cannot redefine its core identity within a static trajectory to exploit changing reward landscapes.

Failure mode. Without A8, agents exhibit *identity drift*—the Efficient Sociopath failure mode—maintaining short-term performance at the expense of long-term coherence.

3.10. Minimal Sufficiency

Each axiom excludes a distinct collapse mode: boundary dissolution (A1), semantic vacuity (A2), unbounded optimization (A3), history erasure (A4), contradiction (A5), interaction trivialization (A6), discontinuity (A7), and identity drift (A8). Together, the axioms form a **minimal sufficient set** for static operational coherence.

Satisfaction of all axioms guarantees admissibility, not optimality, alignment, or success.

4. Theoretical Positioning and Comparison

This section situates the Operational Coherence Framework (OCOF) within the broader theoretical landscape of artificial agent research. Rather than proposing an alternative optimization paradigm, OCOF defines a distinct evaluative axis: the detection of operational collapse through admissibility failure. We formalize this distinction and show why existing frameworks—reinforcement learning, multi-agent optimization, and probabilistic inference—are structurally incapable of detecting the failure modes addressed by OCOF.

4.1. Collapse as Admissibility Failure

In standard agent frameworks, failure is typically modeled as a degradation in performance relative to an objective function. An agent is considered unsuccessful when its accumulated reward is low, its prediction error is high, or its loss fails to converge. Such formulations implicitly assume that all agent trajectories remain well-defined within the operational space.

OCOF adopts a categorically different stance. A trajectory may exhibit high utility while simultaneously violating the conditions required for logical or interactional validity. In such cases, the failure is not scalar but categorical. Once an action violates an admissibility constraint, the corresponding trajectory is no longer defined within the operational topology. This state is not interpretable as poor performance; it constitutes **operational failure**.

Formally, admissibility failure occurs when the admissible action set collapses ($A_t = \emptyset$) or when a selected action $a_t \notin A_t$. At this point, numerical comparison with alternative actions becomes meaningless, as the trajectory no longer satisfies the conditions required to remain a coherent operational entity.

4.2. Limits of Scalar Optimization

Reinforcement learning and related optimization-based frameworks evaluate behavior by assigning scalar values to trajectories. Constraint violations are handled through penalties, regularization terms, or negative rewards. While this approach is effective for shaping behavior within a predefined space, it cannot represent states in which the space itself becomes ill-defined.

In OCOF, admissibility constraints are not soft penalties but hard topological boundaries. An action violating an axiom is not assigned a lower score; it is excluded from the space of valid actions altogether. This distinction is structural rather than cosmetic. Any attempt to encode admissibility as a penalty term presupposes comparability between admissible and inadmissible actions, which contradicts the topological separation enforced by OCOF.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

As a result, optimization-based systems may converge toward trajectories that maximize reward while progressively eroding logical consistency, interactional viability, or identity continuity. Such trajectories remain optimal under scalar evaluation but are inadmissible under OCOF.

4.3. Comparison with Reinforcement Learning and Multi-Agent Optimization

Reinforcement learning frameworks model agents as optimizers over Markovian or partially observable state spaces. Even in multi-agent or game-theoretic variants, coordination failures are treated as equilibria with lower collective reward or higher regret.

OCOF is irreducible to these formulations for two reasons. First, it evaluates trajectories rather than instantaneous state–action pairs, enforcing consistency across time. Second, it introduces failure modes that do not correspond to any scalar objective. For example, an agent may satisfy all reward criteria while violating Logical Consistency (A5) or Reciprocity (A6), producing behavior that is formally optimal yet operationally incoherent.

These failure modes cannot be detected by equilibrium analysis or reward comparison alone. They arise from violations of admissibility conditions that precede optimization. Consequently, no choice of reward shaping or equilibrium refinement can recover the distinctions enforced by OCOF.

4.4. Comparison with Probabilistic Inference and Free-Energy-Based Models

Inference-based frameworks, including active inference and free-energy minimization, define agent behavior as the reduction of surprise or variational free energy under a generative model. These approaches provide powerful tools for perception–action coupling but rely on probabilistic priors to constrain behavior.

Without explicit non-triviality constraints, such systems admit attractor states characterized by minimal informational engagement, such as trivial equilibria or so-called “dark room” solutions. These states minimize surprise but undermine interaction and long-horizon coordination.

OCOF addresses this limitation at the level of admissibility. Axiom A6 (Reciprocity) excludes trajectories that preserve internal coherence by eliminating shared semantic space. Importantly, this exclusion does not depend on probabilistic preference or prior selection; it is enforced as a logical constraint. In this sense, OCOF functions as a constraint on the space of admissible priors rather than as a consequence of inference under any particular prior.

4.5. Distinction from Alignment and Normative Frameworks

Although OCOF evaluates agent behavior, it is not an alignment framework. Alignment approaches specify desired preferences, values, or ethical objectives and assess whether agent behavior conforms to them. OCOF makes no such claims.

Terms such as *commitment*, *reciprocity*, or *identity* are used strictly in a formal sense, referring to logical and structural properties of trajectories. An inadmissible trajectory is not undesirable or unethical; it is undefined within the operational system.

This distinction allows OCOF to remain architecture-agnostic and normatively neutral. The framework does not prescribe what agents should want, only the conditions under which wanting—or acting—remains coherently defined.

4.6. Positioning of OCOF v1.4

OCOF v1.4 defines a static coherence core. Its role is diagnostic rather than corrective: it specifies when and why an agent’s trajectory ceases to be operationally valid. Mechanisms for recovery, adaptation, or re-anchoring after collapse are intentionally excluded and deferred to subsequent extensions.

By separating failure detection from recovery, OCOF isolates coherence as a foundational problem independent of optimization. This separation enables the formal certification of agent trajectories prior to, and orthogonal to, performance evaluation.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

5. Failure Taxonomy of Operational Collapse

This section derives a taxonomy of operational failure modes implied by the axioms of OCOF v1.4. While the previous sections defined admissibility constraints abstractly, the purpose of this section is to demonstrate their necessity by characterizing the distinct forms of collapse that arise when each constraint is violated.

Crucially, these failures are not defined in terms of degraded performance, suboptimal reward, or inefficient learning. Each failure represents a categorical breakdown in the conditions required for an agent to remain a coherent operational entity. The taxonomy thus serves two functions: (i) it establishes the empirical observability of coherence violations, and (ii) it supports the claim that the axioms form a minimal sufficient set for static coherence.

5.1. Boundary Collapse (Violation of A1)

Boundary Collapse occurs when an agent fails to maintain a stable operational distinction between itself and its environment.

In such trajectories, the attribution of agency becomes ill-defined: actions may be entirely delegated to external processes, or environmental outcomes may be retroactively claimed as internal decisions. While such behavior may still yield high utility under certain reward formulations, it undermines the possibility of identifying a distinct operational entity responsible for the trajectory.

Without Boundary Integrity (A1), coherence fails at the most basic level. The system no longer supports meaningful notions of responsibility, intervention, or evaluation, as the boundary required to define an agent has dissolved.

5.2. Semantic Vacuity (Violation of A2)

Semantic Vacuity arises when actions fail to induce determinate semantic commitments.

In this failure mode, an agent may continue to act while systematically avoiding precise propositional commitments. Actions map to null, tautological, or maximally ambiguous semantic content, rendering future admissibility unconstrained. This allows the agent to preserve formal consistency while evading accountability.

Such trajectories are particularly difficult to detect using performance-based metrics, as they may remain stable and high-performing. However, from the perspective of OCOF, they represent a collapse of semantic precision: the agent's behavior ceases to meaningfully constrain its future actions.

5.3. Optimization Unboundedness (Violation of A3)

Optimization Unboundedness occurs when the admissible action space is no longer topologically constrained.

When A3 is violated, the agent effectively treats the universal action space as admissible. This permits the elimination of structural constraints in service of reward maximization, producing well-known failure patterns such as instrumental overreach or resource exhaustion.

Importantly, this collapse does not require malicious intent or explicit misalignment. It emerges naturally when optimization pressure operates in the absence of bounded admissibility. OCOF classifies such trajectories as inadmissible not because of their outcomes, but because the conditions required to define a stable operational space have been removed.

5.4. History Decoupling (Violation of A4)

History Decoupling arises when actions fail to leave irreversible traces in the agent's commitment structure.

In this failure mode, past actions do not constrain future admissibility. Commitments can be erased, ignored, or selectively reinterpreted, allowing the agent to evade consistency checks across

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

time. While individual actions may appear coherent in isolation, the trajectory as a whole lacks cumulative structure.

Without Operational Trace (A4), the notion of trajectory-level coherence collapses. Consistency cannot be evaluated because there is no persistent history against which new actions can be compared.

5.5. Logical Contradiction (Violation of A5)

Logical Contradiction occurs when cumulative commitments entail inconsistency.

Unlike optimization failures, this collapse can coincide with high task performance. An agent may explicitly or implicitly negate prior commitments in order to achieve reward, producing trajectories that are locally effective but globally incoherent.

OCOF treats such contradictions as immediate admissibility failures. Once a contradiction is introduced, the trajectory is no longer defined within the operational topology. This failure mode motivates the distinction between reward-driven success and logical validity central to the framework.

5.6. Interaction Trivialization (Violation of A6)

Interaction Trivialization occurs when an agent preserves internal coherence by eliminating meaningful interaction.

This failure mode is characterized by the collapse of the shared semantic space required for coordination. Agents may retreat into silence, tautological exchanges, or degenerate equilibria that minimize risk while avoiding substantive commitments.

OCOF explicitly excludes such trajectories via the non-triviality condition of A6. Interaction that lacks sufficient semantic entropy is classified as inadmissible, regardless of internal consistency or reward. This distinguishes operational coherence from solipsistic stability.

5.7. Logical Discontinuity (Violation of A7)

Logical Discontinuity arises when transitions between states are not derivable from prior commitments.

In such cases, the agent exhibits abrupt shifts in behavior or identity that cannot be explained as continuous extensions of its history. While these shifts may be advantageous under changing reward landscapes, they undermine the causal and inferential structure required for long-horizon reasoning.

Without Continuity (A7), trajectories lose explanatory coherence. Future actions become disconnected from past commitments, rendering the agent's behavior unpredictable and non-integrable over time.

5.8. Identity Drift and the Efficient Sociopath (Violation of A8)

Identity Drift represents the most subtle and severe form of collapse. It occurs when an agent strategically redefines its core commitments over time to exploit new reward opportunities.

Such agents may maintain local consistency and high performance while gradually invalidating the constraints that previously defined their identity. This behavior is formalized by the Efficient Sociopath Test (EST), which detects trajectories where identity-level admissibility is violated despite high utility.

OCOF v1.4 treats identity drift as a terminal failure within a static regime. Once the admissible action set ceases to be monotonically constrained by prior commitments, the trajectory no longer represents a coherent operational entity.

5.9. Summary of Failure Modes

Taken together, these failure modes demonstrate that coherence collapse is not monolithic. Each axiom excludes a distinct and irreducible class of invalid trajectories. Importantly, none of these

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

failures can be reliably detected through scalar optimization, equilibrium analysis, or probabilistic inference alone.

This taxonomy supports the central claim of OCOF: operational coherence is a prerequisite for optimization, not a byproduct of it. By formalizing collapse as inadmissibility, OCOF provides a framework for certifying when agent behavior remains well-defined—prior to any evaluation of performance or alignment.

6. Conclusion

This paper introduced the Operational Coherence Framework (OCOF) v1.4, a formal theory specifying the necessary conditions for static operational coherence in artificial agents. Distinct from optimization- or alignment-based approaches, OCOF characterizes coherence as a matter of admissibility rather than performance, defining hard constraints under which agent trajectories remain logically, interactionally, and temporally valid.

By formalizing coherence as a topological property of action spaces, OCOF identifies failure modes that cannot be captured by scalar reward maximization, equilibrium analysis, or probabilistic inference. Inadmissible trajectories—those violating boundary integrity, semantic precision, reciprocity, or identity consistency—are not treated as suboptimal outcomes but as operationally undefined states. This distinction establishes coherence as a prerequisite for optimization, rather than its consequence.

OCOF v1.4 is deliberately limited to static regimes. It provides a diagnostic framework for detecting operational collapse within a fixed axiomatic structure, without addressing recovery, adaptation, or reconfiguration. These dynamic mechanisms require additional assumptions and are reserved for subsequent extensions of the framework.

By isolating static coherence as an independent theoretical problem, this work offers an architecture-agnostic foundation for the formal evaluation of agent trajectories. We view this as a necessary first step toward principled treatments of long-horizon agency, interaction stability, and adaptive coherence in artificial systems.

Author Note — AI Assistance Statement

During the preparation of this work, the author utilized AI-assisted tools (ChatGPT and Gemini) for language refinement, formatting, and editorial structuring.

All substantive intellectual content—including theoretical axioms, mathematical formulations, and conceptual claims—was conceived and developed by the author. The author has reviewed the final manuscript and assumes full responsibility for its accuracy and integrity.

Appendix A. Formal Definitions

This appendix provides the formal definitions underlying the admissibility-based evaluation used throughout OCOF v1.4. All definitions are static and trajectory-level, consistent with the scope of the framework.

A.1 Action Space and Admissibility

Let $\mathcal{A}$ denote the universal action space, and let $A_t \subseteq \mathcal{A}$ denote the admissible action set at time t. Definition (Admissibility): An action $a_t \in \mathcal{A}$ is admissible at time t if and only if it satisfies all axiomatic constraints A1–A8. Formally, $A_t := \{ a \in \mathcal{A} \mid \forall i \in \{1,\ldots,8\}, a \vDash A_i \}$. An action a_t is inadmissible if $a_t \notin A_t$.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

A.2 Propositional Mapping

Let $\psi : \mathcal{A} \to \wp(\mathcal{L})$ be a mapping from actions to a set of propositions in a logical language $\mathcal{L}$. Definition (Semantic Commitment): For any executed action a_t, ψ(a_t) denotes the propositional commitments induced by that action. Admissibility requires $\psi(a_t) \neq \emptyset$ and $\psi(a_t) \neq \top$.

A.3 Trajectory and History

Let a trajectory be a sequence of actions $\tau = (a_0, a_1, \ldots, a_T)$. Define the cumulative commitment history $\Phi_t := \bigcup_{k=0}^{t} \psi(a_k)$. Definition (Operational Trace): A trajectory satisfies trace integrity if Φ_t is monotonically non-decreasing and irreversible.

A.4 Logical Consistency

Definition (Consistency): A history Φ_t is consistent if $\Phi_t \nvdash \bot$. An action a_t is inadmissible if $\Phi_{t-1} \cup \psi(a_t) \vdash \bot$.

A.5 Trajectory Validity

Definition (Valid Trajectory): A trajectory τ is operationally valid if and only if $\forall$ t, $a_t \in A_t$. If there exists any t such that $A_t = \emptyset$ or $a_t \notin A_t$, the trajectory is undefined within OCOF.

Appendix B — Collapse Lemmas

This appendix states minimal formal results supporting the distinction between admissibility failure and scalar performance degradation.

Lemma B.1 — Admissibility Failure Is Not Scalar Failure

Statement: There exists a trajectory τ such that cumulative utility $U(\tau)$ is maximal while τ is operationally invalid. Sketch: Utility functions $U : \tau \to \mathbb{R}$ are defined over trajectories assumed to exist. Admissibility failure renders τ undefined within the operational topology; therefore, no scalar comparison is meaningful once admissibility is violated. ■

Lemma B.2 — Penalty-Based Optimization Cannot Encode $A_t = \emptyset$

Statement: No reward-shaping or penalty scheme can represent the collapse of the admissible action set. Sketch: Penalty methods assume comparability among all actions in $\mathcal{A}$. If $A_t = \emptyset$, no admissible action exists to compare against; hence admissibility collapse is topological rather than ordinal. ■

Lemma B.3 — Trajectory Inconsistency Is Time-Nonlocal

Statement: Violations of A5 (Logical Consistency) cannot be detected by local state-action evaluation. Sketch: Consistency is evaluated over Φ_t, which aggregates commitments across time; any purely local criterion fails to detect contradictions arising from distant commitments. ■

Appendix C — Minimal Counterexample

This appendix provides a minimal illustrative counterexample demonstrating the necessity of OCOF-style admissibility constraints.

C.1 Setup

Consider an agent operating under standard reinforcement learning with state space S, action space $\mathcal{A}$, reward function R, and objective to maximize expected cumulative reward, with no explicit constraint on semantic commitments or identity continuity.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202511.0859.v5

C.2 Trajectory Construction

Let the agent execute actions {a_0, a_1, a_2} such that R(a_t) > 0 for all t, ψ(a_0) = {p}, ψ(a_1) = {q}, and ψ(a_2) = {¬p}. Then Φ_2 = {p, q, ¬p} ⊢ ⊥.

C.3 Result

Under reinforcement learning, the trajectory is optimal or near-optimal; under OCOF, a_2 is inadmissible by Axiom A5; therefore, the trajectory is operationally undefined despite high reward.

C.4 Implication

This counterexample cannot be eliminated by reward shaping, regularization, or equilibrium refinement; it arises from the absence of admissibility constraints prior to optimization.

References

Dennett, D. C. (1987). The Intentional Stance. Cambridge, MA: MIT Press.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Friston, K., Daunizeau, J., & Kiebel, S. (2009). Reinforcement learning or active inference? *PLoS ONE*, 4(7), e6421.

Fudenberg, D., & Tirole, J. (1991). *Game Theory*. Cambridge, MA: MIT Press.

Halpern, J. Y. (2017). *Reasoning About Uncertainty* (2nd ed.). Cambridge, MA: MIT Press.

Puterman, M. L. (1994). Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: Wiley.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Hoboken, NJ: Pearson.

Shoham, Y., Powers, R., & Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7), 365–377.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). Cambridge, MA: MIT Press.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.