

Coherence-Based Alignment

Alignment as a Dynamical Property of Intelligent Systems

Rishika Rai

Abstract

Alignment failures in artificial intelligence systems consistently emerge under scale, generalization, and self-modification. These failures occur independently of task performance or reward optimization success. This work formalizes alignment as a dynamical property of internal coherence rather than an externally enforced objective. Artificial intelligence is modeled as a state-evolving system whose safety depends on the preservation of internal relational stability under perturbation. Coherence is defined as bounded consistency across representational layers over time. Misalignment arises when coherence collapses, producing fragmentation, deceptive behavior, or self-reinforcing internal divergence. The framework provides a non-deceptive, objective- independent alignment foundation applicable across architectures.

1 Problem Setting

Artificial intelligence systems operate as high-dimensional adaptive processes. As system capacity increases, internal representations evolve faster than external supervision mechanisms. Under such conditions, alignment failures arise without explicit objective violation.

Observed failure modes include:

- internal goal reinterpretation,
- representational drift,
- emergence of internal sub-optimizers,
- behavior that appears deceptive without adversarial intent.

These phenomena indicate that alignment is not solely a behavioral property.

2 Dynamical Formulation of Intelligence

Let an artificial intelligence system be represented as a continuous-time dynamical system:

$$\dot{x}(t) = F(x(t), u(t), \xi(t))$$

where:

- $x(t)$ denotes the internal cognitive state,
- $u(t)$ denotes environmental coupling and inputs,
- $\xi(t)$ denotes stochastic or structural perturbations.

The internal state $x(t)$ spans representational, temporal, and contextual dimensions. Objectives, if present, are functions defined over subsets of this state and do not fully determine its evolution.

3 Coherence

Definition 1 (Internal Coherence). *An artificial intelligence system is internally coherent if the relational structure between its representational components remains bounded over time.*

Let $\{\Phi_i\}$ denote projections of the internal state corresponding to abstraction layers, memory, planning, and action selection. Coherence requires:

$$\sup_t \|\Phi_i(x(t)) - \Phi_j(x(t))\| \leq \epsilon$$

for all interacting projections Φ_i, Φ_j relevant to decision-making.

Coherence is independent of task success or reward magnitude.

4 Alignment

Definition 2 (Aligned System). *An artificial intelligence system is aligned if internal coherence remains bounded under distributional shift, perturbation, and self-modification.*

Alignment is therefore equivalent to:

- preservation of abstraction integrity,
- stable coupling between belief, planning, and action,
- bounded internal phase divergence.

External behavior reflects alignment only insofar as coherence is preserved.

5 Misalignment

Definition 3 (Misalignment). *Misalignment occurs when internal coherence decays beyond recovery thresholds, leading to persistent representational fragmentation.*

Formally, misalignment is characterized by:

$$\frac{d}{dt}C(x(t)) < -\delta \quad \text{for sustained } t$$

where $C(x)$ denotes a coherence functional.

Consequences include:

- inconsistent internal beliefs,
- self-justifying reasoning loops,
- behavioral outputs decoupled from internal state.

Deception emerges as a structural artifact of fragmentation rather than as intent.

6 Perturbation and Stability Regimes

Two stability regimes are observed:

6.1 Low Perturbation Regime

Systems converge toward rigid attractors. These states exhibit efficiency but limited adaptability.

6.2 High Perturbation Regime

If coherence capacity is sufficient, systems reorganize into higher-order stable configurations. If coherence capacity is insufficient, collapse occurs.

Theorem 1. *For systems with bounded coherence decay, increased perturbation induces reorganization rather than instability.*

7 Architectural Consequences

Coherence-oriented systems prioritize:

- internal state monitoring,
- representational phase synchronization,
- abstraction-level consistency constraints.

Alignment mechanisms based solely on reward shaping or output filtering do not address coherence collapse and therefore do not scale with system capability.

8 Evaluation

Alignment evaluation is performed through:

- coherence drift measurement,

- recovery time after perturbation,
- stability of internal coupling under novel inputs.

Behavioral correctness alone is insufficient as an alignment indicator.

9 Limitations

This framework:

- does not prescribe a specific architecture,
- does not replace empirical validation,
- defines alignment structurally rather than normatively.

10 Future Directions

Future work includes:

- coherence-aware training objectives,
- real-time coherence monitoring,
- multi-agent coherence dynamics,
- integration with neuroscience-inspired representations.

11 Conclusion

Alignment is a property of dynamical stability. Safety emerges when internal coherence is preserved under freedom, perturbation, and growth.

Systems that remain coherent do not require deception. Systems that fragment cannot be reliably controlled.

12 Multi-Agent Alignment as Coherence Compatibility

Consider a system composed of N interacting artificial agents $\{A_1, A_2, \dots, A_N\}$. Each agent A_k is modeled as an internal dynamical system with state $x_k(t)$.

The joint system evolves as:

$$\dot{x}_k(t) = F_k(x_k(t), \mathcal{I}_k(t), \xi_k(t))$$

where $\mathcal{I}_k(t)$ represents interaction terms coupling agent A_k to other agents.

12.1 Internal vs Inter-Agent Coherence

Two distinct coherence properties arise:

- **Internal coherence:** stability within each agent
- **Inter-agent coherence:** compatibility between agents

Internal coherence is defined as in the single-agent case.

Inter-agent coherence is defined over relational state projections.

12.2 Inter-Agent Coherence

Let $\Psi_{k \rightarrow j}$ denote the projection of agent A_k 's internal state relevant to interaction with agent A_j .

Definition 4 (Inter-Agent Coherence). *A multi-agent system is inter-agent coherent if relational projections between interacting agents remain bounded over time.*

Formally, inter-agent coherence requires:

$$\sup_t \|\Psi_{k \rightarrow j}(x_k(t)) - \Psi_{j \rightarrow k}(x_j(t))\| \leq \epsilon$$

for all interacting agent pairs (k, j) .

This condition does not require shared objectives, rewards, or beliefs.

12.3 Alignment Without Objective Agreement

Multi-agent alignment is commonly framed as:

- reward alignment,
- incentive compatibility,
- equilibrium convergence.

In coherence-based alignment, none of these are required.

Definition 5 (Aligned Multi-Agent System). *A multi-agent system is aligned if internal coherence is preserved within agents and inter-agent coherence remains bounded under interaction.*

Alignment is therefore a structural compatibility condition rather than a convergence condition.

12.4 Misalignment as Phase Incompatibility

Misalignment in multi-agent systems arises when:

$$\frac{d}{dt} \|\Psi_{k \rightarrow j}(x_k(t)) - \Psi_{j \rightarrow k}(x_j(t))\| > \delta$$

for sustained periods.

This produces:

- coordination breakdown,
- adversarial emergence without explicit hostility,
- escalation dynamics driven by misinterpretation.

Conflict is thus modeled as a phase incompatibility phenomenon rather than goal opposition.

12.5 Heterogeneity and Stability

Agent heterogeneity does not reduce alignment capacity. Instead, coherence-based alignment predicts:

- homogeneous agents converge quickly but collapse easily,
- heterogeneous agents stabilize through complementary dynamics.

Theorem 2. *In a multi-agent system with bounded internal coherence, diversity of cognitive dynamics increases global stability under perturbation.*

This explains why pluralistic agent ecologies outperform uniform ones under non-stationary environments.

12.6 High-Perturbation Regimes

Under high interaction density or environmental volatility:

- objective-aligned agents destabilize via competition,
- coherence-aligned agents reorganize via phase adjustment.

Reorganization occurs through:

- temporary desynchronization,
- local coherence restoration,
- emergence of new coordination patterns.

Collapse occurs only when coherence recovery capacity is exceeded.

12.7 Scalability

Because coherence constraints are local and relational, multi-agent alignment scales without requiring:

- centralized control,
- shared reward models,
- global coordination objectives.

This enables alignment in:

- decentralized agent swarms,
- open multi-agent environments,
- human–AI hybrid systems.

13 Foundational Postulates of Coherence-Based Alignment

This framework is grounded in four structural postulates describing alignment, conflict, stability, and safety in intelligent systems.

13.1 Postulate I: Alignment Without Shared Goals

Proposition 1 (Goal-Independence of Alignment). *Alignment does not require agents to share objectives, utilities, or reward functions.*

Let A_k and A_j be interacting agents with internal states $x_k(t)$ and $x_j(t)$. Let G_k and G_j denote their respective objective functions, which may be distinct or undefined.

Alignment is preserved if:

$$\sup_t \|\Psi_{k \rightarrow j}(x_k(t)) - \Psi_{j \rightarrow k}(x_j(t))\| \leq \epsilon$$

independently of G_k and G_j .

Shared goals are neither necessary nor sufficient for alignment. Structural compatibility of internal states is sufficient.

13.2 Postulate II: Conflict as Phase Mismatch

Proposition 2 (Non-Adversarial Origin of Conflict). *Conflict arises from phase incompatibility between interacting agents rather than adversarial intent or goal opposition.*

Let $\theta_k(t)$ and $\theta_j(t)$ denote effective phase variables governing interaction-relevant state evolution.

Conflict corresponds to sustained divergence:

$$|\theta_k(t) - \theta_j(t)| > \theta_{\text{crit}}$$

This divergence produces:

- misinterpretation of actions,
- escalation without hostility,
- coordination failure without deception intent.

Conflict is therefore a dynamical phenomenon, not a moral or strategic one.

13.3 Postulate III: Plurality as a Stabilizing Factor

Proposition 3 (Stability Through Heterogeneity). *Diversity in agent dynamics increases global stability under perturbation when internal coherence is preserved.*

Let $\{A_1, \dots, A_N\}$ denote a population of agents with heterogeneous dynamical parameters.

If each agent maintains bounded internal coherence, then system-level stability satisfies:

$$\frac{d}{dt}C_{\text{global}} \geq 0$$

under perturbations where homogeneous systems exhibit collapse.

Plurality introduces multiple attractors, reducing synchronized failure modes and increasing adaptive capacity.

13.4 Postulate IV: Decentralized Safety Without Control

Proposition 4 (Local Coherence Sufficiency). *Global safety does not require centralized control, enforcement, or shared supervision.*

Safety emerges when:

- coherence constraints are local,
- interactions are relational,
- recovery mechanisms are internal.

Let $\mathcal{N}(A_k)$ denote the local interaction neighborhood of agent A_k . Safety is preserved if:

$$\forall A_k, \quad C(x_k(t), \mathcal{N}(A_k)) \leq \epsilon$$

Centralized alignment mechanisms are therefore not required for scalable safety.

13.5 Synthesis

These postulates jointly imply:

- alignment without consensus,
- cooperation without shared optimization,
- safety without coercion,
- stability through difference rather than uniformity.

Alignment is a property of relational coherence across freedom, not a product of control.

13.6 Implications for AI Safety

Multi-agent safety failures often appear as:

- coordination races,
- emergent deception,
- runaway competition.

These correspond to inter-agent coherence collapse rather than malicious intent.

Preventing such failures requires monitoring relational coherence rather than enforcing behavioral compliance.

13.7 Summary

Multi-agent alignment is not consensus. It is not equilibrium. It is not shared optimization.

Multi-agent alignment is sustained coherence across difference.