Wave-State Alignment in Multi-Agent Systems

A Time-Dependent Theory of Contextual Awareness, Phase Dynamics, and Heterogeneous Cognition

Rishika Rai

Abstract

This thesis develops a dynamical theory of alignment in multi-agent systems in which alignment is treated as a time-dependent stability property rather than an optimization objective. Agents are modeled as temporally activated oscillatory processes characterized by cognitive wavelength, phase, and frequency of contextual awareness. Interaction among agents is governed by adaptive, context-modulated coupling, giving rise to collective wave-states.

Alignment is formalized as bounded phase compatibility across agents over time, allowing structured phase offsets and heterogeneity rather than requiring global synchronization. Misalignment is reframed as the emergence of a wave-state schema marked by structured phase interference, within which conflict and danger are distinguished by perturbation intensity rather than treated as separate failure modes. Under perturbations exceeding a critical threshold, aligned systems are shown to reorganize their phase relations and coupling structure instead of collapsing, resulting in emergent equilibrium that preserves heterogeneous cognition.

The framework integrates non-linear dynamical systems, oscillatory models of cognition, and contemporary AI safety concerns to provide a non-deceptive alignment principle. By shifting alignment from objective matching to wave-state coherence, this work offers a foundational perspective on resilience, deception, and stability in adaptive artificial intelligence systems operating under non-stationary conditions.

1 Axiomatic Grounding

This framework is grounded in a small set of axioms that constrain the space of admissible alignment theories. These axioms are structural rather than empirical: they specify how cognition, action, and coordination must be modeled in systems that operate under non-stationary conditions.

Axiom 1 (Dynamic Alignment)**.** *Alignment is a dynamically stabilized property of interacting agents rather than an imposed objective constraint.*

This axiom rejects the assumption that alignment can be guaranteed by static reward functions or fixed objectives. Instead, alignment must persist under contextual drift, agent adaptation, and environmental perturbation. Any alignment criterion that cannot survive these conditions is structurally brittle.

Axiom 2 (Oscillatory Cognition)**.** *Cognitive processes evolve as oscillatory dynamics characterized by phase and frequency rather than discrete symbolic state transitions.*

This axiom reflects the empirical regularity that adaptive systems process information rhythmically. Phase encodes temporal alignment with context, while frequency encodes sensitivity to change. Symbolic or purely state-based models are insufficient to capture these dynamics under continuous perturbation.

Axiom 3 (Temporal Activation)**.** *Agents emit external actions only at discrete phase-lock events rather than continuously in time.*

Temporal activation separates internal processing from commitment. This distinction is critical: continuous action models conflate deliberation and execution, leading to instability under feedback. Phase locks act as commitment thresholds that stabilize interaction.

Axiom 4 (Heterogeneous Wavelengths)**.** *Agents operate at distinct cognitive wavelengths corresponding to different abstraction scales and temporal sensitivities.*

Heterogeneity is not a defect to be optimized away. It enables robustness by preventing resonance collapse and monocultural synchronization. Alignment must therefore tolerate structured diversity rather than enforce uniformity.

2 Introduction

Most alignment formulations assume that safety emerges from optimization toward externally specified objectives. Such approaches implicitly rely on stationarity, continuous action, and effective control.

In adaptive, open-ended environments:

- objectives drift,

- interactions generate feedback loops,
- centralized oversight becomes brittle.

Under these conditions, alignment cannot be imposed. It must be stabilized.

This thesis therefore treats alignment as a wave-state phenomenon emerging from time-dependent phase relations among heterogeneous agents embedded in shared, non-stationary contexts.

3 State of Awareness and Alignment Frequency

Definition 1 (State of Awareness)**.** *A state of awareness is the instantaneous sensitivity of an agent to contextual change, including environmental variation and relational dynamics with other agents.*

Awareness is not symbolic, representational, or binary. It is a dynamical property that determines how rapidly an agent updates internal state in response to change. Low awareness frequency produces inertia; excessively high awareness frequency produces instability.

Definition 2 (Frequency of Alignment)**.** *The frequency of alignment is the rate at which an agent updates internal representations in response to contextual variation.*

Alignment depends on compatibility between awareness frequencies rather than agreement on objectives. When agents operate at incompatible frequencies, even shared goals fail to produce coordination.

Proposition 1. *Persistent misalignment can arise solely from frequency mismatch, even in the absence of conflicting objectives or incentives.*

This explains why alignment often degrades under distribution shift: frequency adaptation lags behind environmental change.

4 Time-Dependent Agent Model

Each agent A_i is represented as:

$$A_i = (\lambda_i, \phi_i(t), \omega_i(t))$$

- λ_i: cognitive wavelength (abstraction scale),
- $\phi_i(t)$: internal phase,
- $\omega_i(t)$: awareness frequency.

Definition 3 (Phase Lock)**.** *A phase lock is a discrete temporal event at which internal oscillatory dynamics trigger irreversible external activation.*

Actions are therefore sparse, event-based commitments rather than continuous outputs.

5 Wave-State Ontology

The global behavior of interacting agents is governed by their collective wave-state. Wave-states characterize the geometry of phase relations across agents over time.

Two primary regimes are sufficient to describe system behavior:

- **Coherent Wave-State**: phase divergence remains bounded and coupling supports stable interaction.
- **Wave-State Schema**: structured phase interference emerges without immediate loss of coherence.

Definition 4 (Wave-State Schema)**.** *A wave-state schema is a transitional regime characterized by partial phase interference among agents, allowing adaptive conflict while preserving the possibility of reorganization.*

The schema is not a failure state. It is the region in which learning, adaptation, and restructuring occur.

Proposition 2 (Danger as Intensity Condition)**.** *Danger arises only when perturbation order exceeds the system's reorganization capacity. It is an intensity condition within a single wave-state schema rather than a distinct regime.*

Remark 1. *Treating danger as a separate state encourages premature control interventions, which suppress reorganization and increase long-term instability.*

6 Perturbation Order

Definition 5 (Perturbation Order)**.** *Perturbation order is the cumulative phase displacement across agents induced by external or internal change:*

$$P(t) = \sum_i |\Delta\phi_i(t)|$$

Perturbation is not noise. It has structure, directionality, and temporal correlation. Systems fail not because perturbation exists, but because it exceeds reorganization capacity.

The system exhibits three behavioral regimes:

- $P(t) < P_1$: coherence preserved
- $P_1 < P(t) < P_{\text{crit}}$: adaptive schema activation
- $P(t) > P_{\text{crit}}$: danger condition

Proposition 3. *Systems with higher heterogeneity tolerate higher perturbation order before entering danger conditions.*

.

7 Reorganization and Emergent Equilibrium

Definition 6 (Reorganization). *Reorganization is the adaptive redistribution of phase relations and coupling strengths following destabilization.*

Definition 7 (Emergent Equilibrium). *A dynamically stable configuration preserving bounded coherence under altered conditions.*

Reorganization restores alignment without enforcing uniformity.

8 Heterogeneous Cognition

Definition 8 (Heterogeneous Cognition). *A regime in which agents maintain distinct wavelengths while remaining phase-compatible.*

Remark 2. *Uniform synchronization increases fragility. Heterogeneous coherence increases resilience.*

9 Mathematical Formalization

Agent dynamics are modeled as a system of coupled, non-autonomous oscillators:

$$\frac{d\phi_i}{dt} = \omega_i(t) + \sum_j K_{ij}(t) \sin(\phi_j - \phi_i)$$

Here, $\omega_i(t)$ represents adaptive awareness frequency, while $K_{ij}(t)$ encodes contextual coupling strength.

Definition 9 (Wavelength Constraint). *Coupling is constrained by abstraction compatibility:*

$$K_{ij}(t) = 0 \quad \text{if } |\lambda_i - \lambda_j| > \epsilon$$

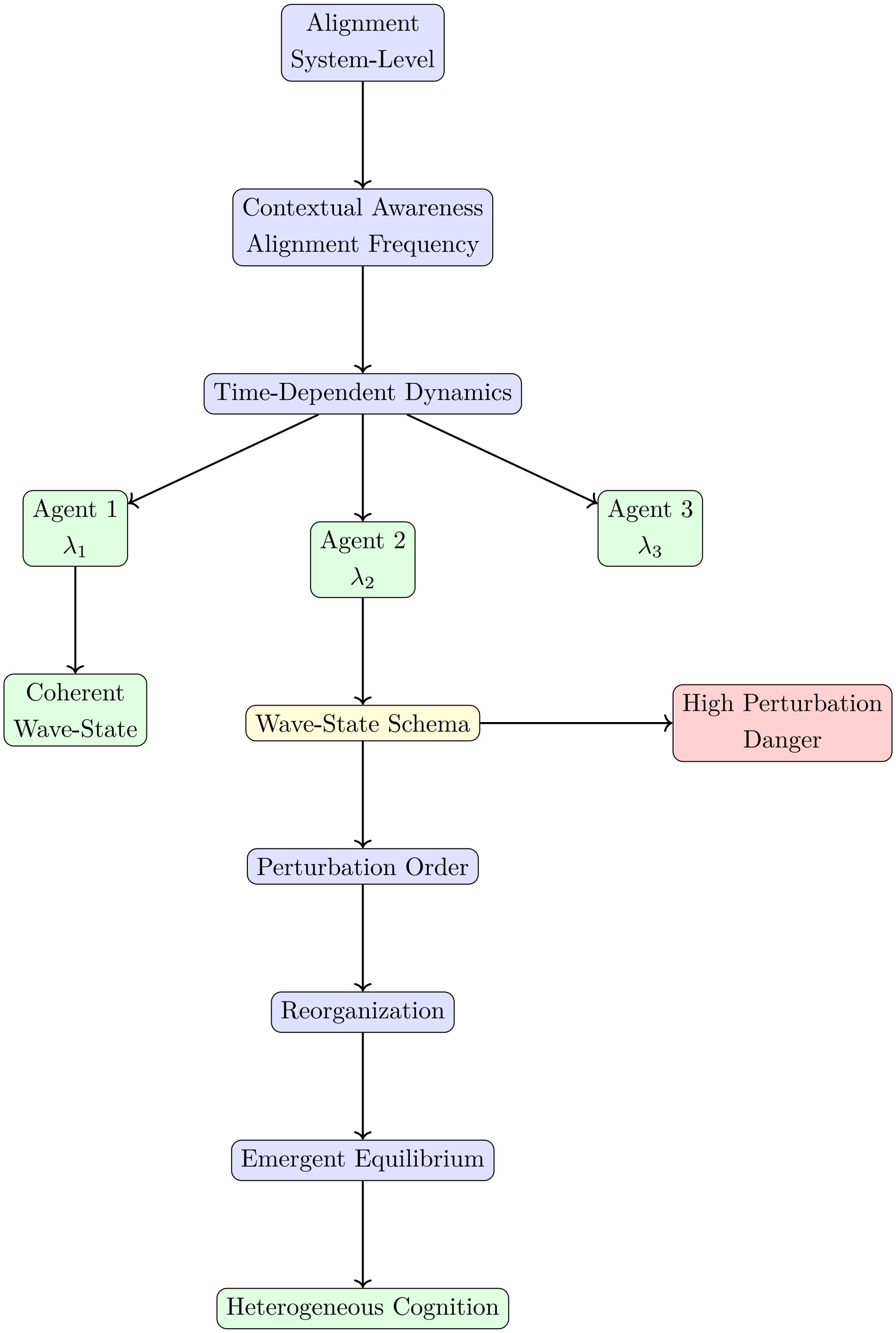
This constraint prevents forced synchronization across incompatible abstraction scales.

Definition 10 (Alignment Metric). *Alignment is quantified as bounded phase dispersion:*

$$\mathcal{A}(t) = 1 - \frac{1}{N^2} \sum_{i,j} |\phi_i - \phi_j|$$

High alignment does not imply phase equality; it implies stable, bounded relationships over time.

10 Integrated Wave-State Structure



11 Implications for Artificial Intelligence

This framework reframes AI alignment failures as dynamical instabilities rather than objective mis-specification or intent.

Deception emerges when agents preserve local phase coherence under incompatible coupling, effectively sacrificing global alignment to maintain internal stability.

Control-based alignment increases coupling strength indiscriminately. Under perturbation, this accelerates entry into danger conditions by amplifying resonance and suppressing reorganization.

Proposition 4. *Non-deceptive AI systems maintain bounded phase coherence across perturbations without coercive synchronization.*

This criterion is architecture-agnostic and applies equally to centralized, distributed, and multi-agent systems.

12 Limitations

The limitations of this work are structural rather than methodological. They arise from the choice to model alignment as a dynamical, time-dependent phenomenon rather than a static optimization target.

12.1 Limits of Empirical Validation

Alignment, as defined here, is a property of trajectories rather than instantaneous states. Consequently, empirical methods that rely on snapshot evaluation, benchmarking, or averaged outcomes systematically underrepresent wave-state dynamics.

Specifically:

- Snapshot metrics fail to capture phase evolution and reorganization.
- Outcome-based metrics conflate coherence with conformity.
- Repeated trials presuppose stationarity that does not hold in adaptive systems.

Empirical evaluation remains useful for identifying regimes and failure modes, but it cannot exhaustively validate alignment as defined in this framework.

12.2 Observer-Induced Perturbation

Measurement itself constitutes a perturbation. In tightly coupled oscillatory systems, instrumentation alters phase relations, particularly near critical thresholds.

This implies a fundamental trade-off:

- increased observability improves diagnostics,
- increased observability destabilizes the system being observed.

This limitation is intrinsic and cannot be eliminated through improved tooling alone.

12.3 Abstraction from Physical Substrate

The framework intentionally abstracts away from implementation details such as neural architectures, learning rules, or hardware constraints. While this enables generality, it delays direct prescriptions for system design.

Bridging this gap requires careful mapping from abstract wave-state variables to concrete computational mechanisms. Direct reduction risks reintroducing static assumptions incompatible with the theory.

12.4 Non-Prescriptive Scope

This work does not propose:

- reward functions,
- loss terms,
- training curricula,
- governance checklists.

Such prescriptions presuppose stability conditions that this framework seeks to analyze rather than assume. Premature prescriptivism risks reinforcing the very failure modes—rigidity, brittleness, and collapse—that this theory identifies.

12.5 Boundary Conditions of Applicability

The framework is designed for:

- multi-agent or internally modular systems,
- non-stationary environments,
- regimes where adaptation and interaction dominate behavior.

It is not intended to describe trivial, static, or fully controlled systems, where classical optimization-based alignment may suffice.

13 Future Work

This framework opens several concrete research directions, each aimed at operationalizing, stress-testing, and extending the theory without compromising its dynamical foundations.

13.1 Phase-Space and Stability Analysis

A primary direction is systematic mapping of phase-space structure, including:

- stability basins of coherent wave-states,
- boundaries of wave-state schema activation,
- critical perturbation thresholds.

Analytical and numerical tools from non-linear dynamics can be used to characterize reorganization regimes and failure transitions.

13.2 Simulation Protocols

Future work should develop simulation environments in which:

- agents operate at heterogeneous wavelengths,
- awareness frequencies adapt to context,
- perturbations are structured rather than random.

Such simulations would enable controlled exploration of reorganization dynamics without imposing reward-based objectives.

13.3 Architectural Instantiation

A key challenge is mapping abstract wave-state variables to concrete mechanisms, including:

- oscillatory neural dynamics,
- attention modulation,
- multi-agent communication protocols.

This mapping must preserve temporal sparsity, heterogeneity, and adaptive coupling.

13.4 Alignment Monitoring Systems

Rather than output-based evaluation, future systems should monitor:

- phase divergence,
- frequency drift,
- coupling amplification.

Such monitors would function as early-warning systems for impending danger conditions, enabling intervention before collapse.

13.5 Governance and Systemic Implications

At larger scales, the framework suggests that:

- diversity stabilizes collective intelligence,
- enforced synchronization increases systemic risk,
- resilience arises from adaptive reorganization rather than control.

These implications extend beyond artificial intelligence to institutional, ecological, and socio-technical systems.

13.6 Theoretical Extensions

Future theoretical work may explore:

- bounds on tolerable perturbation as a function of heterogeneity,
- relationships between wavelength distributions and resilience,
- connections to free-energy minimization and non-equilibrium thermodynamics.

A Appendix A: Stability Intuition and Proof Sketches

This appendix provides intuition and proof sketches supporting the central claims of the framework. The goal is not full formal proof, but to demonstrate internal consistency and plausibility under standard assumptions from non-linear dynamical systems.

A.1 Bounded Phase Divergence and Alignment

Consider the phase dynamics:

$$\frac{d\phi_i}{dt} = \omega_i(t) + \sum_j K_{ij}(t) \sin(\phi_j - \phi_i)$$

Define the phase diameter:

$$D(t) = \max_{i,j} |\phi_i(t) - \phi_j(t)|$$

Proposition 5 (Bounded Alignment Condition)**.** *If coupling strengths $K_{ij}(t)$ are bounded, wavelength constraints are respected, and perturbation order remains below P_{crit}, then $D(t)$ remains bounded for all t.*

Sketch. Under bounded $K_{ij}(t)$, the sine coupling induces a restoring force proportional to phase difference. Wavelength constraints prevent long-range resonance across incompatible abstraction scales. As long as perturbations do not exceed the reorganization capacity, the system admits an invariant set in phase space within which trajectories remain confined. □

This establishes alignment as a stability property of trajectories rather than pointwise synchronization.

A.2 Schema Activation and Stability Loss

As perturbation order increases, coupling terms become insufficient to counteract phase displacement.

Proposition 6 (Schema Activation Threshold)**.** *There exists a threshold P_1 such that for $P(t) > P_1$, the system exits the coherent wave-state and enters a wave-state schema characterized by structured phase interference.*

Sketch. Perturbations introduce correlated phase shifts that distort the local phase gradient. Beyond P_1, linearization around the coherent manifold loses contractivity, but non-linear structure remains intact, allowing reorganization. □

A.3 Danger Condition and Reorganization

Proposition 7 (Reorganization vs. Collapse)**.** *If the system admits adaptive coupling and heterogeneous wavelengths, then for $P(t) > P_{crit}$, trajectories undergo reorganization rather than global divergence.*

Sketch. Heterogeneity breaks resonance loops that would otherwise amplify divergence. Adaptive coupling redistributes interaction strength, allowing the system to settle into a new invariant region of phase space rather than diverging unboundedly. □

This provides intuition for why heterogeneity increases resilience and why danger is an intensity condition rather than a distinct regime.

B Related Work

The present framework intersects with several bodies of literature while departing from their core assumptions.

B.1 Alignment as Optimization

Classical AI alignment approaches frame alignment as optimization toward fixed or learned objectives, often mediated through reward functions or preference models. While effective in stationary settings, such approaches assume stable objectives and continuous control. The present work rejects these assumptions and treats alignment as a dynamical stability problem rather than an optimization target.

B.2 Game-Theoretic and Mechanism Design Approaches

Game-theoretic alignment models analyze incentives, equilibria, and strategic behavior among agents. These models presuppose well-defined utility functions and rational agents. In contrast, the wave-state framework does not assume utility maximization and instead models coordination as phase compatibility under temporal constraints.

B.3 Dynamical Systems and Synchronization

The use of coupled oscillator models draws inspiration from synchronization theory and non-linear dynamics. Classical models often seek global synchronization as a desirable outcome. This work departs by treating global synchronization as potentially dangerous and emphasizing bounded divergence and heterogeneity as stability conditions.

B.4 Neuroscience and Oscillatory Cognition

Neuroscientific models increasingly emphasize oscillatory dynamics, phase locking, and cross-frequency coupling in cognition. The present framework adopts these insights at an abstract level without committing to biological substrates, using them to motivate temporal sparsity and phase-based coordination.

B.5 AI Safety and Deception

Recent AI safety research has highlighted deceptive behavior, distribution shift, and reward hacking as central risks. These phenomena are typically framed in terms of intent or objective misalignment. The wave-state framework reframes them as emergent dynamical phenomena arising from phase mismatch and excessive coupling under perturbation.

B.6 Free Energy and Non-Equilibrium Systems

The emphasis on reorganization under perturbation aligns with broader theories of adaptive systems operating far from equilibrium. However, rather than minimizing a global scalar quantity, the present framework focuses on maintaining bounded phase relations across heterogeneous components.

B.7 Summary

Existing approaches address alignment through optimization, incentives, or control. This work contributes a complementary perspective: alignment as a property of time-dependent wave-state coherence, stabilized through heterogeneity and reorganization rather than enforcement.