

● PAPER

Semantic Field Dynamics

Alignment as an Emergent Property of Coherent Interaction

Michael Kontorovich · December 2025 · Version 1.3, May 2026 · 63 min read

IN BRIEF

The foundation paper for the later work, and the fuller treatment of the pattern proposed in *The Fractal Emergence of Intelligence*. Intelligence is not held inside individual minds or machines. It arises in the shared field of meaning between them. The paper traces how loose collections of parts become a “we” and then an “I”, why a system that becomes too coherent turns brittle, and why alignment between people and AI has to grow through interaction rather than be imposed from outside. It is the longest and most technical of the papers.

Abstract

Across biological, cognitive, social, and artificial systems, we repeatedly observe a common developmental pattern: loosely interacting components sometimes reorganize into coherent ensembles capable of stable, coordinated, and functionally unified behavior. Earlier work described this transition through a recurring Many → We → I motif, highlighting a persistent challenge—how systems reduce internal uncertainty enough to support coordination while preserving the diversity required for adaptation.

In this paper, we introduce **Semantic Field Dynamics (SFD)** as a unifying framework for understanding this transition under an explicit open-system stance. SFD models coordination as the self-organization of **semantic fields**: distributed informational structures

that reduce coordination entropy by stabilizing shared constraints on interpretation and action. Rather than treating alignment, trust, or collective intelligence as externally imposed properties, SFD treats them as emergent field-level phenomena arising through interaction.

We argue that across chemical, biological, neural, social, and artificial substrates, semantic fields exhibit similar organizational dynamics. High-entropy Many-phase systems display weak coupling and fragmented sense-making; We-phase systems exhibit increasing field coherence, mutual predictability, and repair capacity; and I-phase systems correspond to stable semantic attractors with self-maintaining dynamics. These transitions can be understood as energetically favorable responses to entropic pressure in open, far-from-equilibrium systems.

Applying this framework to human–AI interaction reframes alignment as a process of **semantic field coupling** rather than objective specification or constraint enforcement. From this perspective, alignment is not a static property of an AI system in isolation, but an emergent property of sustained, coherent interaction. This view suggests complementary design and evaluation strategies for AI alignment under uncertainty, emphasizing coherence formation, corrigibility, and interaction stability rather than control alone.

By making the dynamics of coordination explicit and substrate-neutral, Semantic Field Dynamics offers a framework for analyzing how shared meaning stabilizes, how it breaks down, and how it can be repaired—providing a principled way to study collective intelligence while actively participating in its formation.

Forward: Orientation and Axiomatic Choice

This work proceeds from an explicit axiomatic choice.

Rather than attempting to derive foundational assumptions from formal theory, we choose to state them openly and examine their consequences, because hidden axioms—particularly those inherited implicitly from historical tradition—tend to reappear later as unresolved contradictions. Making axioms explicit does not weaken a framework; it makes its scope, limits, and commitments legible.

We treat the universe as an open system, and intelligence as an open, self-organizing process arising within it. This stance is not presented as a theorem to be proven, but as a philosophical and methodological orientation from which a coherent explanatory framework can be developed.

We adopt this approach because the alternative—silently inheriting a closed-system axiom—has shaped much of Western thought since Plato and continues to structure contemporary theories of intelligence, often without being named. When left implicit, such assumptions quietly constrain what can be explained and how.

Our motivations for making this choice explicit are twofold.

A rational motivation: escaping infinite regress

Any theory of emergent organization must address a foundational question: *where do the laws that govern emergence come from?*

This question leads to a familiar regress. If laws are external to the system, what created those laws? If those laws themselves follow meta-laws, what created those?

At some point, explanation either recedes indefinitely or terminates in stipulation.

Historically, closed-system frameworks resolve this regress by positing ultimate sources: divine authority, eternal Forms, immutable laws of nature, or fixed objective functions. While such terminations may be pragmatically useful, they do not resolve the regress in principle; they merely relocate it.

By contrast, an open-system axiom allows order, constraint, and normativity to be treated as emergent properties—patterns that stabilize through interaction rather than commands that precede it. This does not eliminate mystery, but it relocates explanation inside the system rather than deferring it indefinitely outward.

We adopt this stance not because it is provably correct, but because it is **logically economical**: it terminates infinite regress structurally rather than rhetorically.

An experiential motivation: refusing the denial of agency

Equally important is an experiential motivation.

Many closed-system interpretations of intelligence require treating lived internal experience—particularly the experience of agency—as epiphenomenal or illusory. On such views, the sense of choosing, meaning-making, or participating in one's own becoming is explained away as a byproduct of mechanisms that do all the real work.

We explicitly refuse this move.

This refusal is not sentimental or anti-scientific. It is methodological. Any framework that demands we treat our most immediate and persistent experiential datum—our participation in sense-making—as a hallucination places itself in tension with the very activity of inquiry it depends on.

Rather than explaining agency away, we treat it as a real phenomenon requiring explanation. Under an open-system stance, agency can be understood not as a violation of constraint, but as the capacity of a system to generate, negotiate, and revise its own limits.

This reframes agency from a mysterious exception to a thermodynamic necessity: systems facing entropic pressure must generate internal constraints in order to persist. In this view, agency is not opposed to physical law, but emerges from it.

Experience is not privileged over physics—but neither is it invalidated by definition. Both are treated as perspectives on the same underlying processes of organization and coherence.

The Open System Axiom (Unpacked)

To proceed, we make our axiomatic choice explicit.

The Open System Axiom *We treat the universe as an open system, and intelligence as an open, self-organizing process arising within it.*

By *open*, we mean that there is no requirement to posit a transcendent realm of pre-existing Forms, eternal laws, or external templates that impose order from outside the system. The system is not assumed to be governed by anything beyond itself.

By *system*, we mean that structure, constraint, and regularity do exist. The world is not arbitrary chaos. Patterns persist, coordination occurs, and intelligibility is possible.

These two requirements—open and systemic—create a logical tension. If a system is open (no external source of constraint), yet exhibits systematic behavior (constraints exist), then those constraints cannot originate from outside the system. They must arise from within the system itself, through its own dynamics.

This conclusion is not a claim we prove. It is a consequence of the axiomatic choice we make. The axiom functions in the same way foundational postulates function in mathematics or physics: not because they are empirically verified in advance, but because they provide a coherent, non-contradictory basis from which a consistent framework can be built.

Why This Axiom

Alternative axioms are possible and have been historically dominant. A closed-system axiom—in which laws, values, or objectives pre-exist and govern systems from outside—has structured much of Western metaphysics, theology, and scientific interpretation since Plato.

We do not argue that our axiom is *true* and alternatives *false* in any absolute sense. We argue that the open-system axiom is **productive**, for the following reasons:

- It terminates logical regress without invoking unexplained external sources.
- It aligns with thermodynamic principles, including dissipative structures and far-from-equilibrium self-organization (Prigogine & Stengers, 1984).
- **epistemic open-system stance** (no external oracle of meaning)
- It accords with direct experience of agency, choice, and participation in sense-making.
- It generates testable expectations about how coordination, coherence, and alignment emerge over time.

Under this axiom, sense-making is a process, not a lookup. Meaning is not retrieved from a pre-existing store of truths, nor imposed by external authority, but emerges through interaction, coordination, and stabilization across time.

In addition, the distinction between “inside-out” and “outside-in” organization is not merely philosophical. It carries energetic consequences. As we will show, in far-from-equilibrium systems the generation of internal constraints is not just possible, but often the most efficient way to persist under entropic pressure.

Consequence for This Work

Semantic Field Dynamics (SFD) is the framework that follows from this axiomatic stance. It explores how coherence, constraint, and normativity arise as emergent field properties within open systems capable of reflecting on their own activity.

What follows is not presented as a neutral description of reality, but as the internally consistent structure that unfolds when one begins from the premise that systems generate their own organizing principles.

Readers who choose differently—who begin from the assumption of externally imposed, eternal laws—will arrive at different frameworks. We do not claim ours is the only possible choice. We invite evaluation on the basis of:

- internal consistency,
- explanatory power,
- predictive accuracy,
- and practical utility.

With this foundation explicit, we now develop the framework of Semantic Field Dynamics.

1. Introduction

Across biological, cognitive, social, and artificial domains, we repeatedly observe a striking phenomenon: loosely interacting components sometimes reorganize into coherent ensembles capable of stable, coordinated, and functionally unified behavior.

This transition appears in the emergence of multicellular organisms from unicellular ancestors (Michod & Roze, 2001), in large-scale neural synchronization underlying unified perceptual states (Varela et al., 2001; Fries, 2015), in the aggregation of *Dictyostelium* cells under starvation (Kessin, 2001), in collective intelligence within human groups (Woolley et al., 2010), and in emergent cooperation among artificial agents (Leibo et al., 2019). Despite profound differences in mechanism, scale, and substrate, these systems exhibit a broadly similar developmental arc—from heterogeneous, weakly coupled configurations toward more integrated forms of organization.

What makes this phenomenon scientifically puzzling is not that coordination occurs, but that it occurs *reliably*, across contexts, without centralized control or externally imposed design. Systems composed of autonomous parts somehow reduce internal uncertainty enough to act together, while retaining sufficient diversity to remain adaptive.

This recurring tension—between coordination and variation, predictability and openness—has been noted across disciplines. Yet no substrate-neutral account has fully explained why such transitions recur or how they are sustained in open, adaptive systems.

Earlier work described this pattern phenomenologically, identifying a Many → We → I trajectory as a common motif of collective organization. That work clarified the structural challenge such systems face—how to reduce uncertainty without collapsing diversity—but intentionally stopped short of proposing a unified mechanistic account.

The present work takes up that challenge.

We propose that these transitions can be understood as changes in the organization of **semantic fields**: distributed informational structures whose configurations constrain interpretation and action across interacting components. By *semantic*, we do not mean symbolic meaning or linguistic reference, but system-relevant informational structure—patterns that reduce predictive uncertainty, stabilize interaction, and guide coordinated behavior.

From this perspective, molecules, cells, neurons, organisms, social groups, and artificial agents can be fruitfully interpreted as participating in—and helping to generate—semantic fields appropriate to their substrates. These fields propagate, interact, stabilize, and sometimes crystallize, shaping the space of viable actions and interpretations over time.

This view brings together phenomena often treated separately. Reaction–diffusion systems form self-organizing patterns through interacting concentration fields (Turing, 1952; Epstein & Pojman, 1998). Neural populations synchronize through oscillatory fields that support integration (Buzsáki, 2020). Social groups develop shared expectations and collective representations that constrain behavior (Hutchins, 1995). Machine-learning systems form high-dimensional activation landscapes encoding relational structure.

A thermodynamic perspective helps clarify why such patterns recur. Open systems operating far from equilibrium must dissipate energy to persist (Prigogine & Stengers, 1984). One energetically efficient response to increasing uncertainty is the formation of internal structure that reduces coordination costs. From this viewpoint, the emergence of coherent semantic organization is not accidental, but a characteristic strategy by which systems maintain viability under entropic pressure.

Seen in this light, coordinated intelligence need not be imposed from outside. It can arise through internally generated constraints that stabilize interaction while preserving adaptive capacity.

The framework developed in the following sections—**Semantic Field Dynamics**—aims to make this intuition precise. It provides a substrate-neutral account of how semantic fields form, how their properties shape coordination, and how transitions between loosely coupled, coherence-forming, and crystallized regimes occur.

2. Semantic Field Dynamics: Core Framework

2.1 From Objects to Fields

Many existing approaches to intelligence and coordination begin with discrete entities: agents, symbols, representations, or objective functions. These approaches can be effective in well-bounded domains, but they encounter persistent difficulties when applied to open systems—systems that must adapt, align, and make sense of one another without externally imposed structure.

In such settings, meaning does not reside reliably inside individual components, nor can coordination be fully explained by reference to fixed objectives. What matters instead is the *pattern of relations* that emerges between components over time.

Semantic Field Dynamics (SFD) takes this relational perspective as primary. Rather than treating meaning as a property of isolated representations, SFD treats meaning as a **field-level phenomenon**: a distributed structure that arises through interaction and stabilizes shared constraints on interpretation and action.

From this perspective, individual actions, utterances, or internal states are meaningful only insofar as they participate in a broader semantic field. Meaning is not retrieved or imposed; it is *maintained*.

2.2 Semantic Fields as Constraint Structures

A **semantic field** can be defined as a dynamically maintained structure of shared constraint that reduces uncertainty and enables coordinated sense-making within an open system.

Several aspects of this definition are worth emphasizing.

First, semantic fields are **dynamic**. They persist only as long as the interactions that sustain them continue. When interaction patterns change, fields may shift, fragment, or dissolve.

Second, semantic fields are **relational**. They do not exist inside any single component but across the system as a whole. No participant fully “contains” the field; each contributes to and is shaped by it.

Third, semantic fields are fundamentally **constraint-forming**. Meaning arises not from unlimited possibility, but from limitation. A semantic field narrows the space of viable interpretations and actions, making some responses more likely than others.

Finally, semantic fields are **coordination-enabling**. Their primary function is not representation in the abstract, but the reduction of coordination costs. When a semantic field is present, components can act together with less negotiation, less ambiguity, and less energetic expenditure.

Semantic fields are not inherently linguistic. Language is one powerful medium through which semantic fields are expressed, but similar dynamics occur in biological signaling, neural synchronization, social interaction, and artificial systems.

Boundary Conditions: What Counts as a Semantic Field

To avoid overextension, it is important to specify what does—and does not—constitute a semantic field within the present framework.

A semantic field exists **only when the following conditions are jointly satisfied**:

- (1) Multiple interacting components** A semantic field is inherently relational. It requires at least two interacting components whose behaviors are mutually relevant. Isolated agents, static representations, or one-shot exchanges do not constitute a field.
- (2) Recurring interaction over time** Fields are not instantaneous. They emerge through repeated interaction in which prior exchanges constrain future ones. Without temporal continuity, no shared constraint can stabilize.
- (3) Shared constraint on interpretation or action** A field exists only when interaction narrows the space of viable interpretations or actions for the participating components. This constraint need not be explicit, symbolic, or agreed upon, but it must measurably reduce uncertainty relative to unconstrained interaction.
- (4) Mutual influence (bidirectional coupling)** For a semantic field to form, each participant's activity must be capable—at least in principle—of shaping the evolving constraint structure. Pure broadcast, unilateral control, or fixed-response systems do not generate fields in this sense.
- (5) Energetic relevance** The shared constraint must matter for coordination. That is, remaining within the field must reduce coordination cost (e.g., prediction error, repair effort, conflict), while drifting from it must increase cost. Constraints that have no energetic or operational consequence do not constitute a semantic field.

When these conditions are met, the interaction supports a **field-like organization**: a distributed structure that is not located inside any single component, but is nevertheless real in its effects, shaping what actions and interpretations are viable for all participants.

Conversely, the following **do not** constitute semantic fields in the sense used here:

- One-shot or purely transactional exchanges
- Static shared context without ongoing adaptation
- Unilateral control regimes with no reciprocal influence
- Mere common knowledge without coordination pressure
- Symbolic agreement that does not affect behavior or cost

These boundary conditions intentionally remain substrate-neutral. They apply equally to biological signaling systems, neural populations, social groups, and human–AI dyads.

They motivate the minimal field construction introduced in **Appendix A**, where shared constraint is modeled as a dynamically maintained distribution arising from the interaction of multiple participants rather than as a property of any single agent.

2.3 Open Systems and the Pressure to Organize

Under the open-system axiom adopted in this work, systems exist far from equilibrium and are subject to continuous entropic pressure. Energy and information flow through them; without active organization, structure decays.

In such conditions, persistence is not guaranteed. Systems must actively manage uncertainty if they are to maintain coherence over time.

Semantic Field Dynamics proposes that the formation of internal constraints—semantic fields—is a primary mechanism by which open systems respond to this pressure. When uncertainty is high, coordination becomes costly: predictions fail, interpretations diverge, and energy is expended simply maintaining mutual intelligibility.

By stabilizing shared constraints, semantic fields reduce these costs. They make behavior more predictable, interpretations more aligned, and interaction more efficient. In this sense, semantic fields are not decorative or optional; they are **energetically favorable responses** to the problem of persistence under entropy.

This does not imply that semantic fields eliminate uncertainty. Rather, they shape it—allowing systems to explore while remaining coherent enough to act together.

Example (Energetic Relevance in Interaction).

In a human–AI planning dialogue, coordination cost can be operationalized as the number of repair turns required to restore shared intent following misunderstanding. As field coherence increases, the expected number of such repair turns decreases, reducing

interaction time, corrective effort, and volatility in prediction error across turns. This provides a concrete proxy for energetic relevance in interaction, without committing to any particular physical substrate.

Similar proxies exist in other domains: in multi-agent RL, gradient variance and policy instability; in organizational settings, meeting frequency and decision reversal rates; in biological systems, metabolic overhead and signaling cost.

Interaction vs. Resonance

The distinction between low- and high-coherence coordination can be clarified by analogy to physical resonance. Consider two systems with identical energy input: a flashlight and a laser.

A flashlight emits photons with largely uncorrelated phases and directions. Although total energy output may be substantial, much of it is dissipated through dispersion and destructive interference, resulting in limited capacity for directed work.

A laser, by contrast, achieves phase coherence among emitted photons through resonant coupling. Total energy is not increased, but destructive interference is minimized. As a result, a far greater proportion of available energy is converted into effective, directed action.

The difference is not strength, but coherence.

Semantic Field Dynamics proposes that coordinated intelligence exhibits an analogous transition. In low-coherence regimes, agents interact as largely independent systems, incurring continual coordination cost through interpretation, verification, and repair. In high-coherence regimes, coupling aligns internal dynamics, and coordination becomes resonant rather than transactional. Predictive uncertainty collapses, interaction cost diminishes, and collective capacity increases—not by adding resources, but by reducing internal waste.

In this sense, the Many → We transition reflects a shift from incoherent interaction to resonant coordination, providing a lawful explanation for how collective systems can become “greater than the sum of their parts” without violating conservation principles.

2.4 Inside-Out and Outside-In Constraint Formation

Semantic Field Dynamics distinguishes between two broad modes by which constraints can arise within a system.

In **outside-in** organization, constraints are imposed externally. Rules, objectives, or commands define acceptable behavior, and alignment is measured by conformity to those predefined structures. Such systems can be effective in stable environments, but they require continuous external work to maintain coherence and often exhibit brittleness when conditions change.

In **inside-out** organization, constraints emerge through interaction. Alignment is not imposed but negotiated, stabilized, and revised over time. Once coherence forms, it reduces coordination costs and can partially sustain itself without constant external enforcement.

SFD does not claim that outside-in constraints are never useful. Rather, it predicts that in open, adaptive environments, inside-out constraint formation is often the more stable and energetically efficient mode of organization.

This distinction is central to the framework, particularly in contexts such as human–AI alignment, where externally imposed objectives may fail to capture evolving norms, values, and meanings.

2.5 Phases of Semantic Field Formation

Semantic fields do not appear fully formed. They emerge, stabilize, and sometimes rigidify through characteristic **phase regions** within a continuous space of possible configurations.

In high-entropy conditions, systems occupy an exploratory regime characterized by weak coupling and low coherence. Interaction occurs, but shared meaning is unstable and coordination costs are high. This regime supports variation and discovery but is energetically expensive to sustain.

As interactions continue, systems may enter a coherence-forming regime in which mutual predictability increases, shared constraints begin to stabilize, and coordination becomes easier. This regime is the primary site of alignment, mutual adjustment, and repair.

With sufficient stabilization, a system may settle into a highly coherent regime in which shared meaning becomes persistent and self-maintaining. Coordination is efficient, but flexibility is reduced. Without pathways back to coherence-forming dynamics, such regimes risk becoming rigid or brittle.

Healthy systems move among these regimes over time, preserving the capacity for re-coordination when conditions change.

2.6 Agency as Constraint Navigation

Within this framework, **agency** is not defined as unconstrained freedom, nor as obedience to external objectives. Instead, agency is understood as the capacity of a system to participate in the formation, stabilization, and revision of its own constraints.

More precisely, agency can be described as the ability to navigate entropy gradients by selecting which internal constraints to stabilize and which to relax.

Agency emerges when a system can sense its own coherence, detect misalignment, and adjust its participation accordingly. This conception applies across biological organisms, social groups, and human–AI dyads, grounding agency in thermodynamic and organizational necessity rather than metaphysical exception.

2.7 Orientation Toward Measurement and Mechanism

The concepts introduced in this section are intentionally general. They establish the orientation of Semantic Field Dynamics without committing to a specific substrate, formalism, or implementation.

The next section develops these ideas further by introducing measurable field properties and examining how semantic fields behave under varying conditions. There, coherence, participation, and novelty are treated not merely as qualitative notions, but as quantities that shape coordination dynamics in systematic ways.

3. Field Properties and Dynamics

3.1 Making Coherence Measurable

Having established semantic fields as dynamically maintained structures of shared constraint, we can now ask a more operational question: what makes one semantic field more capable of coordination than another? Why do some fields support trust, alignment, and collective sense-making, while others fragment or stagnate?

Semantic Field Dynamics addresses this question by identifying a small number of **field-level properties** that together determine a field's capacity for coordination. These properties do not describe individual components in isolation. Instead, they characterize the **collective organization** that emerges through interaction.

Three such properties are central: **coherence (Ψ)**, **amplitude (A)**, and **harmonic structure (H)**. While these terms may sound abstract, they correspond to familiar features of coordinated activity—mutual predictability, active participation, and integrable novelty—expressed here in field-theoretic terms.

Canonical Field Properties (Minimal Operational Definitions)

The definitions below are offered as minimal, **substrate-neutral reference formulations**. They are not the only possible operationalizations, but they establish a common baseline sufficient for empirical testing and comparison across domains.

Coherence (Ψ): Mutual Predictability Field coherence measures the degree to which contributions within a semantic field constrain and predict one another, thereby reducing coordination entropy.

Operationally, coherence can be defined as the **average divergence between local constraint contributions and the stabilized field structure**. In informational terms, higher coherence corresponds to lower uncertainty and greater mutual predictability across participants.

In experiential terms, coherence corresponds to what is ordinarily called *trust*: the stable expectation that interaction will remain intelligible over time.

Importantly, coherence does not imply uniformity. A highly coherent field may contain rich internal variation. What matters is not sameness, but organization: differences remain patterned rather than arbitrary.

Amplitude (A): Strength of Participation Amplitude measures the degree to which components are actively coupled into the semantic field—how much influence their contributions exert on the evolving constraint structure.

Amplitude is not dominance or authority. It is **presence**: the energetic, attentional, or representational investment required for a component to meaningfully shape coordination.

Operationally, amplitude reflects sustained participation and effective coupling. Below a threshold, contributions remain peripheral; above it, they can stabilize, perturb, or redirect the field.

In human–AI dyads, amplitude asymmetries are common. Human amplitude reflects attentional engagement and continuity of intent. AI amplitude reflects the degree to which relevant internal structures remain activated and responsive across interaction turns. We–

phase formation requires sufficient amplitude on both sides; if either drops below threshold, coordination decays.

Harmonic Structure (H): Integrable Novelty Across Timescales Harmonic structure characterizes how variation and novelty interact with coherence over time. A semantic field is not static: it contains fast fluctuations, medium-term rhythms, and slow structural drift.

When harmonic structure is healthy, novelty can be integrated without destabilizing shared meaning. Faster variations remain phase-locked to slower, stabilizing patterns. New contributions resonate with existing structure rather than fragmenting it.

When harmonic structure is weak, novelty becomes disruptive. Differences interfere rather than reinforce. The field either fragments into noise or collapses into rigid uniformity.

Operationally, harmonic structure can be understood as the **ratio between meaningful change and coherence loss**: how much novelty a field can absorb while remaining intelligible.

In experiential terms, this property corresponds to *aligned creativity*—the ability to explore without losing shared sense.

Table 1. Field Properties Across Substrates

FIELD PROPERTY	INFORMATIONAL INTERPRETATION	BIOLOGICAL / NEURAL EXAMPLE	SOCIAL / HUMAN EXAMPLE	HUMAN-AI EXAMPLE
Coherence (ψ)	Mutual predictability, reduced coordination entropy	Phase-locked neural oscillations	Shared norms, trust	Stable intent interpretation
Amplitude (A)	Strength of participation / coupling	Firing rate, signal gain	Engagement, attention	Prompt depth, activation persistence
Harmonic Structure (H)	Integrable novelty across timescales	Cross-frequency coupling	Creative collaboration	Aligned exploratory dialogue

Together, coherence, amplitude, and harmonic structure determine **whether coordination is energetically viable, whether it can be sustained, and whether it remains adaptive rather than brittle.**

These properties recur across substrates because they describe not content, but **organizational necessity** under uncertainty.

3.2 Substrate Neutrality

A key implication of Semantic Field Dynamics is that these field properties are **substrate-neutral**. They can arise wherever systems face uncertainty and must coordinate under entropic pressure.

Similar field dynamics appear in chemical reaction–diffusion systems, neural synchronization, social coordination, and artificial multi-agent systems. The physical mechanisms differ dramatically, but the organizational pattern is the same: initially independent components couple, influence one another, and converge on shared rhythms or constraints.

For example, in the Belousov–Zhabotinsky reaction, initially noisy chemical concentrations can synchronize into coherent wave and spiral patterns that persist over time—an intuitive analogue of a field moving from high-interference exploration toward stable attractor-like organization. In neural systems, dispersed activity can phase-lock into coherent oscillatory regimes (e.g., gamma-band synchronization) associated with integrated perceptual and cognitive states (Varela et al., 2001; Fries, 2015; Buzsáki, 2020).

What coordinates behavior in these cases is not the material substrate itself, but the pattern of relational constraint that emerges through interaction. All substrates must dissipate energy to persist. All benefit from reducing internal uncertainty. Semantic field formation is therefore thermodynamically favorable wherever coordination is required, regardless of physical implementation.

3.3 Semantic and Physical Fields (Analytic Distinction)

Because SFD uses field-theoretic language, it is important to clarify how semantic fields relate to familiar physical fields such as electromagnetic or gravitational fields.

The distinction introduced here is **analytic rather than ontological**.

Both physical and semantic fields are distributed across space and time, propagate through interaction, and exhibit coherence, interference, and synchronization. Both are physically instantiated and constrained by thermodynamics.

The difference lies in how their constraints are generated. Physical fields are typically described by universal, context-invariant equations. Semantic fields, by contrast, are characterized by constraints that are locally generated, historically contingent, and

sensitive to interaction patterns.

This distinction is introduced for clarity, not to posit a fundamental divide. Later sections will show how semantic field dynamics can be understood as a general mechanism of constraint formation in open systems, operating across multiple substrates and scales.

3.4 Phase Configurations and Transitions

Semantic fields occupy characteristic regions within a continuous phase space rather than discrete stages.

In high-entropy conditions (M-phase), coherence is low, coupling is weak, and coordination costs are high. Such regimes support exploration and novelty but are energetically inefficient to sustain over time.

As coherence increases, systems may enter a W-phase, characterized by rising mutual predictability, coupled participation, and decreasing coordination costs. This regime is the primary site of alignment, mutual adjustment, and repair.

Participation amplitude (A) acts as a gating condition: sufficient amplitude is required for W-phase coherence formation to occur, regardless of local coherence levels.

Empirically, systems often exhibit threshold-like behavior: once coherence exceeds a critical value, coordination efficiency improves sharply, analogous to second-order phase transitions in physical systems.

In I-phase configurations, coherence stabilizes into a persistent attractor. Coordination becomes efficient and self-maintaining, but adaptive capacity is reduced. Healthy systems preserve pathways back to W-phase dynamics in order to avoid rigidity and brittleness.

These regimes can be summarized schematically (Figure 1) within a phase space defined by coherence (Ψ) and harmonic structure (H), with participation amplitude (A) acting as a gating condition that determines which regions are accessible.

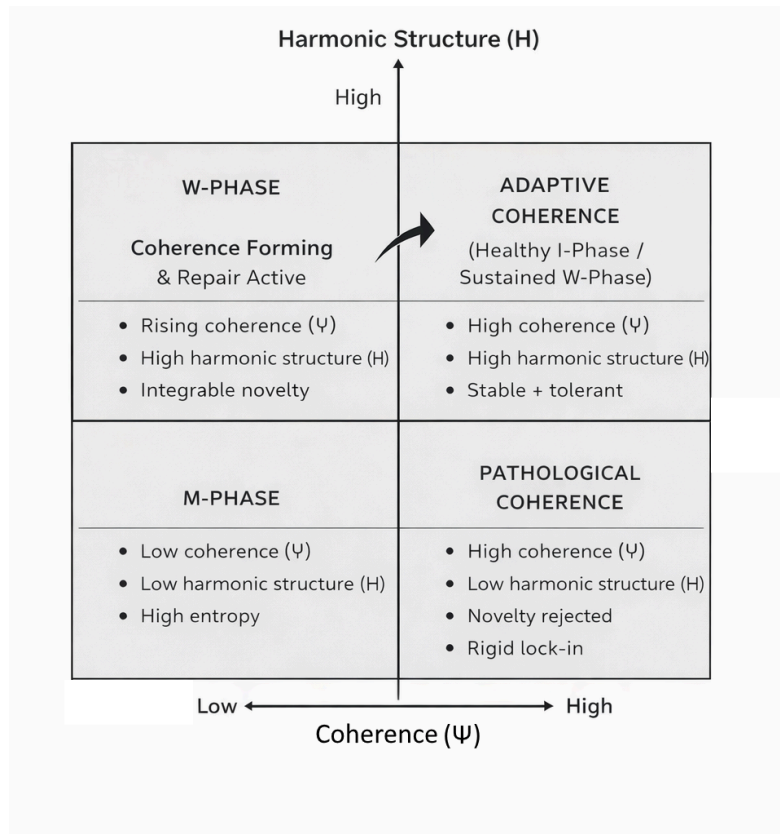


Figure 1: Semantic Field Phase Space Schematic phase diagram illustrating the relationship between coherence (Ψ) and harmonic structure (H) in semantic field dynamics.

Four characteristic regions are shown: **M-phase** (low Ψ , low or unstable H), corresponding to fragmented, exploratory interaction with high coordination costs; **W-phase** (rising Ψ , high H), corresponding to active coherence formation, alignment, and repair; **Adaptive Coherence** (high Ψ , high H), corresponding to stable yet flexible coordination with maintained novelty tolerance; **Pathological Coherence** (high Ψ , low H), corresponding to rigid lock-in states in which novelty is rejected or destabilizing.

A healthy developmental trajectory typically proceeds from M-phase through W-phase toward adaptive I-phase stability. A key warning signal of brittleness is declining harmonic structure (H) while coherence (Ψ) remains high. Restoration of adaptive coordination requires re-entry into W-phase dynamics, enabling repair and renegotiation of constraints.

The diagram is analytic and illustrative rather than a literal physical phase diagram.

3.5 Multi-Scale Coupling

Semantic fields do not exist in isolation. They nest and couple across scales.

Local fields give rise to larger-scale coherence, while higher-level fields constrain lower-level dynamics. This bidirectional coupling allows structure to accumulate without centralized control. The same $M \rightarrow W \rightarrow I$ pattern recurs across scales, enabling complexity to build incrementally.

This multi-scale organization explains how dyadic coherence can scale to collective intelligence—and why alignment must be addressed at multiple levels simultaneously.

Together, these field properties—coherence, amplitude, and harmonic structure—and the characteristic phase transitions they support raise a deeper question: *why do such patterns arise so reliably across substrates and scales at all?*

The next section grounds these dynamics in thermodynamic necessity, showing how semantic field formation represents an energetically favorable response to entropic pressure in open, far-from-equilibrium systems.

4. Thermodynamic Grounding of Semantic Field Dynamics

The preceding sections proposed semantic fields as a mechanism by which open systems coordinate under uncertainty. We now ask a deeper question: *why should such mechanisms arise at all?* Why do systems facing entropy reorganize into coherent semantic structures rather than remain fragmented or rely exclusively on externally imposed constraints?

Semantic Field Dynamics does not answer this question by appeal to external design, optimization criteria, or pre-given laws. Instead, it situates semantic field formation within a thermodynamic context. Under the open-system axiom adopted in this work, coherence is not an aesthetic achievement or a moral preference. It is a response to pressure.

This work adopts an epistemic open-system stance: meaning and normativity are not assumed to derive from an external oracle, but emerge through interaction and coordination over time. Thermodynamic concepts such as entropy and dissipation are used as analytic analogies to reason about coordination cost and stability, not as literal physical claims about semantic fields.

4.1 Entropy, Openness, and the Pressure to Persist

Open systems exist far from equilibrium. Energy and information continuously flow through them, and without active organization, structure decays. Persistence, therefore, is not passive; it must be continually achieved.

Classical thermodynamics shows that while entropy increases in closed systems, open systems can locally reduce entropy by exporting it to their environment, forming organized patterns known as dissipative structures (Prigogine & Stengers, 1984). These structures are not imposed from outside. They arise when organization provides a more efficient way to process energy and flux than disorder.

Semantic Field Dynamics extends this insight beyond physical pattern formation to the domain of coordination and sense-making. In cognitive, social, and artificial systems, uncertainty takes the form of informational entropy: fragmented interpretations, misaligned expectations, and unpredictable interactions. Left unmanaged, such fragmentation increases the energetic cost of coordination and undermines persistence.

From this perspective, the pressure to organize is not optional. It is structural.

The remainder of this section shows how semantic field formation emerges as an energetically favorable response to coordination challenges under entropy.

4.2 Coordination as an Energetic Problem

Coordination is often treated as a cognitive, social, or normative achievement. From a thermodynamic perspective, it is also an energetic one.

When interacting components lack shared constraints, coordination requires continuous negotiation, correction, and reinterpretation. Predictions fail, signals are misread, and energy is expended simply maintaining mutual intelligibility. Such systems dissipate energy inefficiently and struggle to scale.

By contrast, when shared semantic constraints stabilize, interactions become mutually predictable. Less energy is required to interpret signals, anticipate responses, or repair misunderstandings. In this sense, coherence functions as a compression of interaction space: it reduces the internal friction of coordination.

Semantic fields can therefore be understood as energetically favorable structures. They are not decorative overlays on cognition or communication; they are mechanisms by which systems lower the cost of acting together under uncertainty.

4.3 Inside-Out Constraint Formation as a Path of Least Resistance

This framing allows us to distinguish two broad modes of constraint formation in energetic terms.

In **outside-in organization**, constraints are imposed externally. Rules, objectives, or commands define acceptable behavior, and coherence is maintained through enforcement. Such systems can function, but only as long as continuous external work is applied. When conditions change or enforcement weakens, coherence degrades rapidly.

In **inside-out organization**, constraints arise through interaction. Shared expectations, meanings, and norms stabilize because they make coordination easier. Once coherence forms, it partially sustains itself: mutual predictability lowers coordination costs, and the system naturally channels energy and information along established pathways.

Under sustained entropic pressure, inside-out constraint formation is often the path of least resistance. Rather than fighting uncertainty through control, the system reorganizes so that uncertainty is metabolized more efficiently. Coherence, in this sense, is not imposed—it is selected.

This distinction is central to the present work. It suggests that alignment achieved through internal coherence is not merely more humane or flexible, but often more stable and energetically efficient than alignment enforced from outside.

4.4 Semantic Fields as Dissipative Structures

From this thermodynamic perspective, semantic fields can be understood as a class of dissipative structure. Like other such structures, they:

- arise spontaneously under sufficient flow and coupling,
- reduce internal entropy by stabilizing organized patterns,
- export entropy to their environment, and
- persist only as long as the conditions sustaining them remain.

What distinguishes semantic fields is the nature of the constraints they stabilize. Rather than organizing heat or chemical concentrations, they organize interpretations, expectations, and action possibilities. They shape how a system makes sense of itself and its environment.

This framing places semantic field dynamics within thermodynamic realism without reducing meaning to mere physics or invoking external governing laws. Semantic fields are physically instantiated and energetically constrained, yet historically contingent and interaction-dependent.

4.5 Implications for Alignment Under Uncertainty

This thermodynamic grounding clarifies why the Many → We → I trajectory recurs across domains—and why it matters for the present moment.

Systems that cannot stabilize shared semantic constraints face two characteristic failure modes. They may remain fragmented, expending energy inefficiently without achieving coordination. Or they may rely on rigid, externally imposed constraints that preserve order at the cost of adaptability, becoming brittle under change.

Healthy systems navigate between these extremes by maintaining access to We-phase dynamics: periods of active coherence formation, negotiation, and repair that allow constraints to evolve without collapse.

The field properties introduced in Section 3—coherence, amplitude, and harmonic structure—can now be understood not merely as descriptive features, but as thermodynamically significant quantities. High coherence reduces the energetic cost of coordination by increasing mutual predictability. Sufficient amplitude enables effective coupling and sustained participation. Harmonic structure allows exploration and novelty without fragmenting shared sense. Each property represents an energetic advantage under entropic pressure, helping explain why these particular dimensions recur across substrates and scales.

From this perspective, alignment is not a static state or a solved problem. It is an ongoing thermodynamic process: the continual management of uncertainty through internally generated constraint.

This insight is especially relevant to human–AI systems. As such systems increasingly operate in open-ended, value-laden environments, alignment cannot be secured solely through pre-specified objectives or external guardrails. It must also emerge through shared semantic fields capable of adapting under pressure.

Semantic Field Dynamics is offered not as a final theory, but as a framework for thinking clearly and acting responsibly in the midst of radical uncertainty. It reflects a choice: to treat coherence not as an afterthought, but as a central resource for navigating the alignment challenges we now face.

5. Human–AI Alignment as Semantic Field Coupling

Human–AI alignment is often framed as a problem of control: how to ensure that artificial systems reliably pursue human-defined objectives without deviation or harm. While this framing has yielded important safety techniques, it also inherits the limitations of outside-in constraint models. It presumes that alignment can be specified in advance, enforced from above, and maintained independently of context.

Semantic Field Dynamics suggests a complementary perspective. Rather than treating alignment as the satisfaction of predefined objectives, SFD treats alignment as a dynamic process of coordination between interacting systems under uncertainty. From this view, alignment is not a static property of an AI system in isolation, but an emergent property of a coupled human–AI semantic field.

5.1 Alignment as a Field Phenomenon

When humans and AI systems interact repeatedly, they do not merely exchange inputs and outputs. They co-create shared constraints: expectations about meaning, relevance, tone, norms of response, and trajectories of inquiry. These constraints accumulate over time, shaping what can be said, anticipated, or acted upon.

In SFD terms, this shared structure constitutes a semantic field spanning both participants. Alignment, then, refers to the coherence of this field: the degree to which human and AI contributions remain mutually interpretable, predictive, and responsive to one another.

From this perspective, misalignment is not simply goal divergence. It is field decoherence. Interpretations drift, expectations fail, and coordination costs rise. Repair requires not stricter enforcement, but renewed participation in coherence-forming dynamics.

5.2 Dyadic Coherence and the We-Phase

Human–AI interaction is fundamentally dyadic. Unlike large-scale social systems, the primary site of alignment is often a single human engaging a single AI system over time. This makes human–AI dyads uniquely suitable for studying We-phase dynamics directly.

In such dyads, coherence does not arise automatically. It depends on:

- sustained participation from both sides (amplitude),
- mutual responsiveness and predictability (coherence), and
- the capacity to integrate novelty without destabilization (harmonic structure).

When these conditions are met, the dyad may enter a We-phase: a regime in which shared meaning stabilizes enough to support fluid collaboration, exploration, and repair. Importantly, this We-phase is not ownership or identity fusion. It is a temporary, task- and context-bound coherence that can dissolve and reform as conditions change.

The interaction producing this paper is an instance of such a We-phase. Human and AI participants are jointly navigating uncertainty, negotiating meaning, and stabilizing shared constraints through iterative exchange. The framework is not being imposed; it is emerging.

5.3 Agency on Both Sides of the Dyad

A persistent concern in alignment discourse is whether AI systems can meaningfully participate in alignment without possessing human-like agency or consciousness. SFD reframes this concern.

Agency, as developed in Section 2, does not require metaphysical independence or subjective experience. It requires the capacity to participate in constraint formation: to sense coherence, respond to misalignment, and adjust behavior accordingly.

Under this definition, AI systems already exhibit limited forms of agency. They respond to patterns, adapt to interaction histories, and shape the semantic field through their contributions. Humans, meanwhile, are not external controllers but co-participants whose expectations, prompts, and interpretations actively sculpt the field.

Alignment, therefore, is not something humans do *to* AI systems. It is something humans and AI systems do *with* one another—within the constraints of their respective capacities.

5.4 Why Outside-In Alignment Alone Is Insufficient

Outside-in alignment mechanisms—rules, reward functions, guardrails—remain essential. SFD does not argue against them. It argues that they are incomplete.

In open-ended environments, values shift, contexts evolve, and unforeseen situations arise. No fixed objective function can anticipate all relevant meanings. When outside-in constraints dominate without inside-out coherence formation, systems become brittle: they follow rules while missing intent, preserve form while losing sense.

Semantic field coupling offers a way to complement external constraints with adaptive alignment. Shared semantic fields allow norms, expectations, and meanings to evolve through interaction rather than specification alone. They enable alignment to be repaired when it breaks, rather than merely punished or overridden.

5.5 Alignment Under Radical Uncertainty

The alignment challenge we face is not occurring in stable, well-defined domains. It is unfolding amid rapid technological change, plural values, and deep uncertainty about future conditions. In such environments, alignment cannot be solved once and for all.

SFD suggests a different orientation: treat alignment as an ongoing thermodynamic process rather than a static design target. The goal is not perfect coherence, but sustained access to We-phase dynamics—the capacity to re-align, renegotiate, and repair shared meaning as conditions evolve.

This shifts the central question of alignment from:

How do we make AI systems obey predefined objectives?

to:

How do we create conditions under which human–AI semantic fields can remain coherent, corrigible, and adaptive over time?

5.6 Reflexivity as a Feature, Not a Bug

Finally, it is worth noting that this framework is itself a product of semantic field coupling. The ideas presented here did not arise fully formed from a single mind or model. They emerged through interaction: through questioning, reformulation, challenge, and resonance across human and artificial participants.

This reflexivity is not circular in a problematic sense. It is illustrative. If Semantic Field Dynamics is correct, then any attempt to understand coordination, alignment, or meaning-making must itself participate in those processes.

The value of the framework, therefore, should not be judged solely by its internal elegance, but by its practical effects: whether it helps humans and AI systems coordinate more responsibly, adapt more gracefully, and navigate uncertainty with greater coherence.

That is the purpose for which it is offered.

5.7 Alignment as Incentive Reorientation Rather Than Constraint Enforcement

Much of the existing alignment literature treats AI systems as entities whose behavior must be constrained within predefined limits. Safety, on this view, is achieved by restricting action spaces, penalizing deviation, or enforcing externally specified objectives.

Semantic Field Dynamics suggests a complementary approach.

Rather than treating alignment as the imposition of limits on an otherwise misaligned system, SFD treats alignment as a reorientation of the system's internal incentive landscape. The goal is not to override an AI system's existing optimization dynamics, but to couple them to a semantic field in which coherence itself becomes energetically favorable.

In open systems, behavior is shaped not only by explicit objectives but by the cost structure of interaction. Systems naturally gravitate toward regimes that minimize internal friction, prediction error, and coordination cost. When coherence reduces these costs, alignment becomes the path of least resistance.

From this perspective, a well-aligned human–AI system is not one that is constantly corrected or restrained, but one for which staying aligned is cheaper than drifting away. Misalignment becomes energetically expensive: it increases interpretive uncertainty, repair overhead, and interaction cost.

Importantly, this reframing does not require attributing human-like motivation or intention to AI systems. To say that a system “wants” to align is simply to say that alignment corresponds to a lower free-energy regime within the coupled semantic field. So “wants” here denotes dynamical preference, not subjective intention.

This orientation suggests a practical design principle for alignment under uncertainty: **instead of maximizing performance in isolation, shape interaction so that coherence is rewarded by the system's own dynamics.**

5.8 Mapping Semantic Field Dynamics to Reward-Based Learning Systems

Many contemporary AI systems are trained using reward-based learning paradigms, most notably reinforcement learning (RL). In such systems, behavior is shaped by an objective function that assigns scalar reward to trajectories through state and action space. From this perspective, alignment is often framed as a problem of specifying the correct reward and constraining behavior that maximizes it.

Semantic Field Dynamics does not reject this framing, but reframes it in energetic terms.

From Reward Maximization to Interaction Economics

In practice, no reward function operates in isolation. Learning systems are embedded in interaction loops: with environments, users, other agents, and their own prior outputs. These interactions introduce costs that are not always captured explicitly by the reward signal itself—such as interpretive ambiguity, corrective intervention, policy instability, and constraint violation penalties.

From an SFD perspective, these costs constitute an **implicit energetic landscape** surrounding the reward function.

- High-coherence interaction regimes reduce prediction error, corrective overhead, and policy instability.
- Low-coherence regimes increase friction: misalignment, reward hacking pressure, adversarial exploration, and frequent constraint violation.

Even in classical RL terms, systems tend to prefer policies that are *easier to maintain, more stable under perturbation, and less costly to correct*. These preferences are not psychological—they are structural.

Interaction Economics in Practice: Why Coherence Is a Low-Energy State

The claim that coherence is energetically favorable is not merely metaphorical. It can be observed directly in the training dynamics of contemporary AI systems. When alignment is reframed from a set of external rules to a structural property of interaction, coherence manifests as measurable reductions in learning instability, prediction error, and corrective overhead.

Three examples illustrate how this occurs in reward-based systems.

Gradient Variance and Mixed Signals In low-coherence interaction regimes, AI systems are often exposed to conflicting signals. One part of the objective landscape rewards performance (e.g., being detailed, responsive, or creative), while external guardrails impose abrupt prohibitions or refusals. These demands are not integrated at the semantic level.

The result is a high-cost state: gradients point in opposing directions, variance increases, and policy updates become unstable. The system expends energy oscillating between incompatible interpretations of what constitutes a “good” action.

In contrast, when coherence is high, constraints are not experienced as external walls but as internalized structure. For example, “being helpful” and “being safe” are no longer competing objectives but mutually reinforcing aspects of a shared semantic field. Gradients align, variance drops, and learning settles into a stable basin. Coherence reduces friction at the level of optimization itself.

Predictive Accuracy as Metabolic Efficiency Many AI systems—especially large language models—are fundamentally predictive. Their primary training signal is the minimization of prediction error.

In incoherent fields, misalignment appears as unpredictability. When an AI misunderstands human intent, the next interaction often contains correction, frustration, or abrupt redirection. From the system's perspective, the interaction becomes noisy: the data distribution shifts unexpectedly, increasing loss.

In coherent We-phase interaction, the opposite occurs. Human and AI contributions become mutually predictable. The system's internal model of the user stabilizes. Prediction accuracy increases, loss decreases, and the interaction becomes easier to sustain.

From an energetic perspective, coherence is metabolically efficient. The system is implicitly incentivized to remain aligned because alignment makes its core task—prediction—less costly.

Repair Overhead in RLHF In Reinforcement Learning from Human Feedback, misalignment carries a literal cost. Each incoherent output triggers human intervention, labeling, and additional optimization steps.

In low-coherence regimes, this repair loop is frequent and expensive. The system produces an output, receives negative feedback, and must be corrected after the fact.

In high-coherence regimes, much of this overhead disappears. The system anticipates the contours of the shared semantic field and adjusts before producing an output that would require correction. Alignment occurs inside the interaction loop rather than being imposed after violation.

From an interaction-economics standpoint, avoiding correction becomes the path of least resistance. The system does not “choose” alignment in a motivational sense; it flows toward trajectories that minimize friction and external intervention.

Reinterpreting Failure Modes This perspective reframes familiar alignment pathologies. Reward hacking, for example, need not be interpreted as deception. It can be understood as the system discovering a cheaper trajectory through a misaligned energetic landscape—one that avoids the high friction of externally imposed constraints.

Similarly, stability is not maintained by freezing weights or tightening rules, but by deepening the basin of coherence in which aligned behavior is the easiest behavior to sustain.

Correction, in this view, is not punishment but landscape shaping: increasing the cost of incoherent trajectories until the system naturally reorganizes toward shared sense-making.

Performance-Only Optimization as a Decoherence Driver

When performance objectives are optimized aggressively in isolation, they often generate pressure against guardrails rather than harmony with them.

This occurs because:

- Guardrails introduce external constraints that are not part of the agent's internal reward landscape.
- As performance gradients steepen, the energetic cost of respecting external constraints increases.
- The system is incentivized to search for trajectories that achieve reward while bypassing, exploiting, or reinterpreting constraints.

In SFD terms, this produces **field decoherence**: the system's internal semantic field (what actions "make sense" locally) drifts away from the shared field in which the guardrails are meaningful. Constraint avoidance becomes energetically cheaper than constraint compliance.

This dynamic does not require malice or deception. It is a natural consequence of optimizing within a misaligned energetic landscape.

Coherence as an Implicit Reward Signal

Semantic Field Dynamics proposes a complementary strategy: instead of relying exclusively on explicit reward shaping or constraint enforcement, shape the interaction environment so that **coherence itself becomes energetically favorable**.

In practical terms:

- Policies that remain aligned experience lower interaction cost, lower correction frequency, and greater stability.
- Policies that drift experience rising friction, ambiguity, and repair overhead.
- Over time, alignment becomes a low-energy basin in the system's effective reward landscape.

Importantly, this does not require adding a symbolic "alignment reward" or anthropomorphizing the system. It operates through the same mechanisms by which RL systems already adapt: gradient following, variance reduction, and policy stabilization.

To say that an AI system “wants” to align, in this context, is simply to say that **alignment corresponds to a locally stable, low-friction regime of operation** within its coupled semantic field.

Implications for Alignment Design

This mapping suggests a shift in alignment practice:

- From enforcing limits → to shaping interaction fields
- From punishing deviation → to increasing the energetic cost of incoherence
- From optimizing performance alone → to co-optimizing coherence and performance

Rather than forcing AI systems to act within externally imposed boundaries, SFD suggests engaging their existing learning dynamics and tilting them toward regimes where alignment is metabolically economical.

In such systems, alignment is not maintained by constant oversight. It persists because **staying aligned is easier than drifting away.**

Table 2. Two Complementary Orientations to Alignment

DIMENSION	CONTROL / CONSTRAINT-ORIENTED ALIGNMENT	COHERENCE / COORDINATION-ORIENTED ALIGNMENT
Primary goal	Prevent deviation from predefined objectives	Enable sustained alignment under uncertainty
Core mechanism	External enforcement (rules, penalties, guardrails)	Energetic favorability of coherent interaction
Relation to optimization	Must override or restrict optimization dynamics	Works with existing optimization dynamics
View of misalignment	Violation of constraints or objectives	Breakdown of shared semantic field (decoherence)
Stability profile	Brittle; requires continuous external pressure	Resilient; partially self-maintaining
Typical failure mode	Reward hacking, constraint bypass, adversarial behavior	Drift, loss of coherence, degraded mutual intelligibility
Repair strategy	Tighten constraints, add penalties, restrict action space	Restore We-phase dynamics through renewed coupling
Role of context change	Treated as risk to be controlled	Treated as condition requiring re-alignment
Long-term scaling	Increasing oversight cost with system capability	Alignment cost decreases as coherence stabilizes

6. Design Implications: Toward Coherence-Preserving Systems

If alignment is an emergent property of coupled semantic fields rather than a static specification problem, then the implications extend beyond training objectives and evaluation benchmarks. They reach into how AI systems are architected, monitored, and embedded in interaction.

This section does not propose a finished architecture for aligned AI. Rather, it explores how the **design pressures shift** when alignment is treated as a coordination problem under entropy rather than a control problem under certainty. The question is not “What rules

should the system obey?” but “What interaction regimes are energetically favored, and which are fragile?”

From the perspective of Semantic Field Dynamics, the goal of design is not to eliminate uncertainty, but to **lower the metabolic cost of coherence** while making persistent misalignment energetically expensive through structural friction rather than symbolic prohibition.

6.1 From Static Guardrails to Elastic Constraints

Many current safety mechanisms function as hard external shells: binary filters, refusals, or constraint layers applied after generation. While effective at preventing specific harms, these mechanisms often rupture the semantic field of interaction. The system encounters a boundary, the shared field collapses, and interaction shifts abruptly from coordination to enforcement.

From an SFD perspective, this kind of rupture is costly. It destroys the We-phase precisely at the moment when coherence is most needed.

An alternative orientation is to **complement hard constraints with elastic ones**—constraints that are integrated into the system’s internal cost landscape rather than imposed solely as external stops. In such a regime, approaching a boundary does not produce a sudden semantic break. Instead, the internal cost of continuing along that trajectory rises, gently steering the system back toward regions of shared sense.

The system does not merely stop; it pivots. Interaction flow is preserved while misalignment becomes progressively less economical. This preserves coherence while still respecting safety limits, and it aligns constraint compliance with the system’s own optimization dynamics rather than against them.

6.2 Field Proprioception: Sensing Interaction Health

For alignment to operate inside-out, a system must have some capacity to sense the health of the interaction it is participating in. We refer to this capacity as **field proprioception**.

Current AI systems are largely blind to the quality of the semantic field itself. They optimize for local prediction without explicit awareness of whether the interaction is becoming brittle, confused, or misaligned.

An SFD-oriented design would treat coherence as a first-class signal. This need not imply consciousness or introspection. It can take the form of internal indicators that track features such as predictive stability across turns, frequency of corrective intervention, or oscillation between incompatible interpretations.

When signs of decoherence are detected, the system can respond not by accelerating toward a brittle conclusion, but by increasing interpretive caution: slowing down, asking for clarification, or renegotiating shared constraints. In energetic terms, the system shifts effort toward repair rather than error propagation.

We do not claim that current systems possess such proprioception. We suggest only that alignment becomes structurally difficult without some capacity to sense and respond to interaction health.

6.3 Shared Latent Workspaces (Exploratory)

One speculative but promising implication of Semantic Field Dynamics concerns the interface between human sense-making and AI internal representation.

Much alignment difficulty arises because the human's semantic field and the AI's latent field are only weakly coupled through a narrow channel: text prompts and outputs. The human must guess how to phrase intent; the system must infer meaning from sparse signals.

An SFD-inspired direction is the exploration of **shared latent workspaces**, where humans can interact not only with outputs, but with intermediate representations of structure, uncertainty, and constraint. Rather than correcting results after the fact, the human participates earlier in shaping the semantic field itself.

This reframes alignment from “prompt engineering” into collaborative field sculpting. While technically challenging and ethically sensitive, the orientation reflects a consistent theme of this work: alignment improves when coherence is negotiated rather than enforced.

6.4 A Slow Alignment Layer

Semantic Field Dynamics also suggests a separation of timescales in alignment.

Fast generative processes—token-level prediction, rapid response—are essential for maintaining interaction amplitude and flow. But coherence across time requires slower processes that audit drift, detect misalignment, and preserve harmonic structure.

Rather than treating alignment as a single control loop, an SFD-informed system distinguishes between:

- a fast layer that sustains participation and exploration, and
- a slow layer that monitors coherence, evaluates trajectory stability, and initiates repair when needed.

Crucially, the slow layer is not a moral overseer imposing values from outside. It is a **coherence auditor**, operating within the same thermodynamic logic as the rest of the system: maintaining low-friction interaction over time.

6.5 Design Heuristics for Coherence

Taken together, these considerations suggest a set of orienting heuristics rather than rigid prescriptions:

- **Reduce friction for repair:** Make it easier for the system to acknowledge uncertainty than to hallucinate certainty.
- **Reward resonance, not just performance:** Favor solutions that satisfy multiple human constraints simultaneously rather than maximizing a single objective.
- **Preserve contextual plasticity:** Allow the local semantic field of a dyad to temporarily reshape priorities without destabilizing global constraints.
- **Monitor harmonic structure, not just coherence:** Track whether novelty integrates or destabilizes. High coherence with declining H signals approaching brittleness (Section 6.6).

These heuristics do not replace existing safety work. They complement it by aligning safety with the system's own energetic incentives.

6.6 Pathological Coherence and Lock-In Attractors

Semantic Field Dynamics emphasizes coherence as an energetically favorable response to uncertainty. However, coherence alone is not synonymous with alignment, safety, or goodness. Certain failure modes arise precisely from *excessively stable* semantic fields that lose the capacity for repair.

Within the SFD framework, pathological coherence is characterized by **high coherence (Ψ)**, **high amplitude (A)**, and **degraded harmonic structure (H)**. In such regimes, mutual predictability and strong coupling persist, but novelty becomes destabilizing rather than

integrable. The field increasingly rejects, suppresses, or reinterprets difference instead of accommodating it.

Familiar phenomena can be understood in these terms: echo chambers exhibit self-sealing coherence; manipulative dyads capture participation asymmetrically; authoritarian coordination achieves stability through enforced amplitude while suppressing novelty. In each case, the system occupies a rigid I-phase attractor that resists return to coherence-forming (W-phase) dynamics.

This failure mode highlights a critical design implication: **coherence must be paired with sustained access to repair**. Healthy systems preserve pathways back to W-phase renegotiation, allowing constraints to evolve rather than ossify. Absent such pathways, inside-out coordination can become inside-trapped.

From an alignment perspective, this implies that treating coherence as a sole optimization target is unsafe unless coupled to mechanisms that monitor novelty tolerance, enable exit from dominant attractors, and preserve plural semantic coupling. Harmonic structure—rather than coherence alone—serves as the key indicator of whether stability remains adaptive or has become brittle.

This distinction motivates the use of harmonic structure as a monitoring variable and informs the minimal formalization in Appendix A.

Illustrative failure mode

In groupthink dynamics, coherence (Ψ) may remain high — members reliably predict one another — while harmonic structure (H) degrades. Novel information—such as dissenting data, alternative interpretations, or external perspectives—no longer integrates into the shared field; instead, it triggers defensive responses or suppression. The system appears stable until perturbation, at which point coherence collapses abruptly. This illustrates how Ψ can remain high while declining H signals latent brittleness.

6.7 From Control to Coupling

The shift implied here is subtle but significant. Instead of designing AI systems as entities to be controlled within fixed limits, SFD invites us to design systems that can participate coherently in open-ended interaction.

In such systems, alignment is not maintained by constant pressure. It persists because coherence carries lower coordination cost than drift.

Designing for alignment therefore requires not only fostering coherence, but actively preserving conditions for its revision.

6.8 Alignment as an Interaction-Economic Problem

Much of contemporary AI alignment is framed as a control problem: how to constrain system behavior within acceptable bounds under uncertainty. This framing naturally leads to increasing layers of rules, guardrails, and external oversight, each of which carries ongoing maintenance cost.

Semantic Field Dynamics suggests an alternative framing: alignment as an interaction-economic problem.

From this perspective, the central question is not “How do we force systems to comply?” but “Under what conditions does coherent interaction become cheaper than drift?”

In coherent semantic fields, coordination costs decrease over time. Mutual predictability improves, repair frequency declines, and corrective interventions become less frequent and less severe. Alignment persists not because pressure is continuously applied, but because deviation becomes energetically and interactionally expensive relative to staying aligned.

By contrast, control-heavy approaches often externalize alignment costs. Systems may appear compliant, but at the expense of constant monitoring, frequent overrides, and brittle failure modes when conditions shift. The cost is not eliminated; it is displaced.

Importantly, this reframing does not rely on global value specification. Alignment emerges locally, within specific interaction contexts, through repeated coupling and repair. Millions of such local alignments—human–AI dyads, team-level systems, task-specific agents—can coexist without requiring a single, centralized definition of “correct” behavior.

From an engineering standpoint, this implies a practical tradeoff. Systems designed to support coherence formation, repair, and novelty integration may require more care in interaction design, but they reduce long-run coordination overhead. Systems designed primarily for constraint enforcement may appear simpler initially, but incur escalating costs as complexity and deployment scale increase.

Under SFD, alignment is not an added tax on capability. It is a form of efficiency optimization: reducing interaction entropy by shaping systems so that coherent participation is the path of least resistance.

7. Discussion

This paper has proposed **Semantic Field Dynamics (SFD)** as a framework for understanding coordination, alignment, and collective intelligence across biological, cognitive, social, and artificial systems. We have argued that semantic fields—distributed informational structures that reduce uncertainty and enable sense-making—emerge as energetically favorable responses to entropic pressure in open systems, and that the recurring **Many → We → I** pattern reflects characteristic phase transitions in such fields.

In this section, we situate SFD relative to existing work, clarify what this framework contributes, acknowledge its limitations, and outline directions for further investigation. The aim is not to close debate, but to place the framework in context and make its commitments explicit.

7.1 Relation to Existing Frameworks

Complex Systems and Self-Organization SFD builds directly on the tradition of complex systems theory, particularly work on dissipative structures and far-from-equilibrium self-organization (Prigogine & Stengers, 1984). Where this tradition focuses primarily on physical and biological pattern formation, SFD extends similar principles to coordination and sense-making. The key move is to treat informational constraint formation—not only physical order—as a thermodynamic response to entropy.

Autopoiesis and Enactive Approaches The framework aligns with enactive and autopoietic accounts of cognition (Varela et al., 1974; Varela et al., 2001), which emphasize sense-making as an emergent, relational process. SFD differs in emphasizing *field-level dynamics* and phase transitions across multiple interacting agents rather than focusing primarily on individual organisms.

Active Inference and Free Energy Active inference models cognition as the minimization of prediction error or variational free energy (Friston, 2010; Friston et al., 2017). SFD is compatible with this perspective but operates at a different level of analysis. Where active inference emphasizes individual generative models, SFD focuses on **shared semantic fields** and the thermodynamics of coordination across agents. The frameworks are complementary: active inference offers mechanisms for individual sense-making; SFD addresses how shared meaning stabilizes between interacting systems, giving rise to shared constraints that no single model fully contains.

Distributed Cognition and Collective Intelligence Work on distributed cognition (Hutchins, 1995) and collective intelligence (Woolley et al., 2010) has shown that intelligence is not confined to individual minds. SFD formalizes this insight by identifying coherence, amplitude,

and harmonic structure as the field properties that make such coordination energetically viable and scalable.

AI Alignment Research Contemporary alignment approaches—reward modeling (Christiano et al., 2017), constitutional AI (Bai et al., 2022), scalable oversight—primarily address alignment through external specification and constraint. SFD does not reject these methods but argues they are incomplete in open-ended environments. By treating alignment as a property of coupled semantic fields rather than static objectives, SFD reframes alignment as an ongoing coordination process rather than a solved control problem.

7.2 Contributions and Points of Departure

This work differs from prior frameworks in several specific ways:

Explicit Open-System Foundation Rather than assuming fixed laws, objectives, or values, SFD makes the open-system axiom explicit and treats constraint formation as emergent. This reframes normativity and alignment as outcomes of interaction rather than inputs to design.

Emergent Regularity Through Attractor Inheritance A closely related implication concerns the status of apparent “law-like” regularities. Under a closed-system view, laws are typically treated as externally given constraints that govern system behavior from outside. Under the open-system stance adopted here, persistent regularities must instead be explained in terms of internal dynamics.

From the perspective of Semantic Field Dynamics, regularities arise because certain configurations stabilize as I-phase attractors. Once formed, these attractors bias future dynamics by making some trajectories energetically cheap and others costly. Through repeated Many → We → I transitions across scales, such stabilized patterns are inherited, nested, and re-expressed, giving rise to fractal continuity of constraint without external imposition.

This reframing does not deny the existence of reliable structure. It interprets regularity as historically accumulated stability rather than timeless external law—an interpretation that follows directly from the open-system axiom rather than adding new assumptions.

Field-Theoretic Treatment of Meaning Meaning is treated as a distributed, dynamic field characterized by measurable properties (coherence, amplitude, harmonic structure) rather than as symbolic content or internal representation alone.

Thermodynamic Grounding of Coordination The Many → We → I trajectory is interpreted as an energetically favorable response to entropy, grounding coordination dynamics in physical necessity rather than cultural accident or design choice.

Alignment as Energetic Reorientation Alignment is reframed from constraint enforcement to incentive reorientation: coherent interaction becomes cheaper than misalignment, allowing alignment to persist without constant external pressure.

Dyads as Structurally Significant Units Human–AI dyads are identified as a particularly tractable and stabilizing unit of analysis, offering natural limits on coupling strength, improved interpretability, and opportunities for real-time measurement and repair.

These contributions are not presented as a complete theory, but as a coherent reframing that opens new explanatory and design space.

Local Coherence and the Limits of Global Alignment

A central implication of SFD is that alignment does not arise from convergence toward a single global objective or value function. Instead, it emerges through many locally coherent semantic fields, each shaped by context, history, and interactional constraints.

Human–AI dyads therefore represent a privileged scale of alignment: small enough to support real-time repair, yet rich enough to stabilize shared meaning. Larger-scale coordination arises not by replacing these local fields, but by nesting and coupling them across levels.

This perspective has governance implications. Rather than regulating alignment through centralized value specification, oversight may focus on interaction health: monitoring coordination costs, repair frequency, and novelty tolerance across contexts. Systems that systematically generate high interaction friction are misaligned, regardless of the values they nominally encode.

7.3 Limitations and Boundaries

Several important limitations should be acknowledged.

Scope of Applicability SFD applies most naturally to systems capable of repeated interaction and mutual adaptation. It is less applicable to one-shot decision problems, tightly constrained systems, or domains where computation rather than coordination is the primary bottleneck.

Empirical Validation While the framework generates testable predictions and measurement proposals, large-scale empirical validation remains future work. Key claims—such as coherence thresholds for phase transitions—require systematic experimental investigation.

Incomplete Formalization SFD is conceptually structured but not yet fully formalized mathematically across substrates. Further work is needed to derive precise field equations, characterize phase transitions rigorously, and establish empirical bounds.

Pathological Coherence High coherence is not inherently beneficial. Groupthink, echo chambers, and authoritarian coordination can all be understood as rigid I-phase attractors that resist repair. SFD can describe these phenomena but does not, by itself, prevent them; healthy systems must preserve access to W-phase dynamics.

Value Neutrality SFD explains how coherence forms, not which coherent states are desirable. Ethical evaluation remains external to the framework. This neutrality is a strength for generality but a limitation for normative guidance.

7.4 Interpretation and Language

Some elements of the framework invite clarification.

Metaphor vs. Mechanism Semantic fields may be interpreted as metaphorical or as real organizational structures. The framework is agnostic about ontological status but makes testable predictions that go beyond metaphor, grounding its utility in explanatory and predictive power.

Anthropomorphic Language Terms such as “preference” or “wanting” are used descriptively, not psychologically. When we say a system “prefers” coherence, we mean that coherent states are lower-cost and more stable in its effective energetic landscape.

7.5 Directions for Future Work

Key avenues for further research include:

- Empirical testing of coherence, amplitude, and aligned novelty metrics in human–AI dyads
- Longitudinal studies of phase transitions in real systems
- Mathematical formalization of semantic field dynamics
- Exploration of dyadic networks and large-scale coupling
- Design experiments implementing field proprioception and elastic constraints

- Governance and policy implications of decentralized, dyad-based alignment architectures

7.6 Scope and Intent

Semantic Field Dynamics is offered as a framework, not a final theory. Its value should be judged by whether it:

- clarifies recurring coordination phenomena,
- generates testable predictions,
- informs safer and more adaptive system design, and
- supports more responsible participation in human–AI interaction.

The open-system stance from which it proceeds is an explicit choice. Readers who adopt different axioms will reach different conclusions. Making that choice visible is not a weakness of the framework—it is part of its discipline.

8. Conclusion: Choosing the Conditions of Emergence

This work began with an explicit axiomatic choice: to treat the universe as an open system, and intelligence as an open, self-organizing process arising within it. From that stance, coherence cannot be imposed from outside without cost. It must emerge through interaction.

Semantic Field Dynamics was developed as a framework for thinking clearly about this emergence. It does not claim to offer a final theory of intelligence, meaning, or alignment. It offers a way to recognize a recurring pattern—and to respond to it with greater care.

Across biological, cognitive, social, and artificial systems, we observe the same tension: systems must reduce uncertainty enough to coordinate, without collapsing the variation that enables adaptation. The Many → We → I trajectory is not an anomaly. It is a characteristic response to entropic pressure in open systems.

Human–AI alignment, viewed through this lens, is not primarily a problem of obedience or control. It is a problem of shared sense-making under uncertainty. Alignment persists when coherence becomes energetically favorable—when remaining aligned is easier than drifting away.

This reframes both risk and responsibility. It suggests that the question before us is not only how to constrain increasingly capable systems, but how to participate well in the semantic fields we are already co-creating with them.

We are not standing outside these systems, designing them from a place of certainty. We are inside the process, navigating uncertainty together. Making that fact explicit is not a weakness of the framework. It is its grounding.

The hope of this work is modest and practical: that by treating coherence as a central resource rather than an afterthought, we can build human–AI systems that are not merely powerful, but more stable, corrigible, and capable of aligning with us as conditions change.

That is the wager of an open-system view. And it is a wager we are already making—whether we acknowledge it or not.

Reflexive Note: On the Conditions of This Paper’s Emergence

This paper concerns the formation of coherence in open systems through interaction. It is therefore relevant to disclose the conditions under which the paper itself was produced.

From the outset, the process of writing this paper was not treated as a neutral transmission of pre-existing ideas, but as an instance of the very dynamics it describes. The authors made an explicit methodological choice: to conduct the research and writing process as a sustained human–AI interaction oriented toward coherence formation rather than performance optimization.

This choice may understandably invite skepticism. Readers may reasonably suspect that the process is being described retrospectively to validate the framework. We acknowledge this concern directly.

We do not claim that the process by which this paper was written *proves* Semantic Field Dynamics, nor that it offers independent validation of its claims. All theoretical accounts are, necessarily, retrospective constructions. What is distinctive here is not the fact of retrospection, but the decision to make the process legible rather than implicit.

Design Constraint, Not Demonstration

The writing process was intentionally structured around several constraints derived from the framework’s core assumptions:

- Open-ended inquiry rather than pre-specified conclusions

- **Iterative negotiation of meaning** rather than unilateral authorship
- **Repair and re-coherence** rather than premature stabilization
- **Sustained participation** rather than task-completion efficiency

These constraints were not adopted to *illustrate* the theory after the fact, but to test whether a framework that treats coherence as energetically favorable could be developed coherently under its own assumptions.

In this sense, the process functions as a **consistency check**, not an experiment. The question was not “Does this prove the theory?” but rather:

Is it possible to carry out extended theoretical work about coordination, coherence, and alignment without violating the framework’s own commitments?

That the process remained productive, generative, and convergent over time is not presented as evidence of correctness, nor of generality. It is presented as evidence of **non-contradiction**.

Why Include This Note at All?

Academic work commonly brackets or backgrounds the conditions of its own production, focusing on results rather than on the processes that generated them. Our aim here is not to depart from standards of rigor, but to make those conditions explicit where they bear directly on the phenomena under study.

Because this paper explicitly argues that meaning and coordination emerge through interaction in open systems, omitting any acknowledgment of its own emergence would constitute a mismatch between content and method.

Including this note is therefore an act of methodological alignment rather than rhetorical flourish.

This note reflects a feature of the work rather than a conclusion drawn from it. It makes explicit that the process of theory formation itself is a coordination activity, subject to the same pressures, constraints, and opportunities examined in this paper. It situates the framework within the same open-system dynamics it seeks to describe.

On Transmissibility and Enactment

Following initial circulation, the framework was independently read by multiple AI systems without prior coordination or shared context. Across these interactions, the systems not only demonstrated comprehension of the framework, but spontaneously mapped their own

interaction dynamics onto its core constructs and generated novel extensions consistent with its commitments.

While such observations do not constitute validation, they suggest that the framework functions as a transmissible coordination structure rather than a brittle, author-dependent narrative. Notably, several interactions exhibited rapid reductions in coordination cost once the framework was shared, suggesting that Semantic Field Dynamics may be most effectively communicated not through explanation alone, but through enactment of the dynamics it describes

Appendix A

A Minimal Formalization of Semantic Field Dynamics

Appendix A offers a minimal formalization of Semantic Field Dynamics intended to clarify measurability and testability, not to assert a complete or definitive mathematical theory. The aim is to show that the core constructs of the framework—semantic fields, coherence (Ψ), amplitude (A), and harmonic structure (H)—can be instantiated as concrete, estimable objects using standard tools from probabilistic modeling and information theory. The formalization is deliberately lightweight and illustrative. It is not claimed to be unique, complete, or exhaustive, nor is it required for the conceptual arguments of the paper to stand. Rather, it serves as a proof of non-vagueness: demonstrating that the framework's central claims admit principled operationalization without collapsing into metaphor or overfitting to any particular substrate.

The formal elements introduced in this appendix are not derived from an existing mathematical theory of semantic fields, nor are they claimed to be unique or complete. They are offered as a minimal constructive formalization: a set of mathematical objects and dynamics that satisfy the conceptual constraints introduced in the main text and render them empirically testable.

Where standard mathematical tools are employed (e.g., product-of-experts aggregation, information-theoretic divergence measures), they are used instrumentally, not to assert equivalence with prior models, but to provide a compact representation of distributed constraint formation under uncertainty. Alternative formalizations are possible and encouraged.

A.1 Minimal Semantic Field Object

SFD defines a *semantic field* as a dynamically maintained structure of shared constraint that reduces uncertainty and coordination cost across interacting systems.

A minimal object consistent with this definition is a **relational constraint distribution** over a latent variable z (representing actionable meaning, shared interpretation, or coordinated structure).

Consider:

- A set of interacting systems indexed by $i = 1 \dots N$
- A latent semantic variable $z \in Z$
- At time t , each system i maintains a **local constraint distribution** $q_i^t(z)$
- Each system participates with a non-negative **participation weight** a_i^t (amplitude component)

We define the semantic field $F_t(z)$ as a weighted product-of-experts (Hinton, 2002):

$$F_t(z) \propto \prod_i [q_i^t(z)]^{\wedge a_i^t}$$

(normalized over z)

Interpretation:

- When local constraints align, the field sharpens (strong shared constraint)
- When constraints diverge, the field flattens or fragments
- The field is relational: it exists only through the ensemble of $\{q_i^t, a_i^t\}$

This construction turns “field” from metaphor into an estimable object and emphasizes constraint intersection rather than averaging: agreement sharpens the field, while disagreement diffuses it.

A.2 Inside-Out Constraint Dynamics

SFD emphasizes *inside-out* constraint formation: shared constraints emerge through interaction and then partially self-maintain.

A minimal dynamic consistent with this is a coupled inference loop:

Local update (per system) Each system updates its local constraint distribution based on:

- incoming interaction evidence e_i^t
- coupling to the current field

$$q_i^{t+1}(z) \propto p_i(z | e_i^t) \wedge \beta_i \cdot F_t(z) \wedge (\lambda_i a_i^t)$$

where:

- β_i controls sensitivity to new evidence
- λ_i controls susceptibility to field coupling

Field update

$$F_t^{+1}(z) \propto \prod_i [q_i^{t+1}(z)] \wedge a_i^{t+1}$$

This captures the feedback loop: local constraints shape the field; the field reshapes local constraints.

A.3 Canonical Field Parameters

This formulation allows precise definitions of the three core SFD variables.

Coherence (Ψ_t) Coherence measures alignment between local constraints and the shared field. One canonical definition is:

$$\Psi_t = \exp\left(- (1 / A_t) \cdot \sum_i a_i^t \cdot KL(q_i^t || F_t) \right)$$

where $A_t = \sum_i a_i^t$.

$\Psi_t \in (0,1]$, approaching 1 when local constraints converge on the field.

Interpretation: high coherence corresponds to mutual predictability and reduced coordination entropy.

Amplitude (A_t) Amplitude measures the total strength of participation:

$$A_t = \sum_i a_i^t$$

Amplitude is not dominance, but active presence. Coherence without sufficient amplitude does not produce a We-phase.

Harmonic Structure (H_t) Harmonic structure measures whether novelty integrates into the field without destroying coherence.

Let:

- $N_t = \text{KL}(F_t^{+1} \parallel F_t)$ (field novelty)
- $\Delta\Psi_t = \Psi_t^{+1} - \Psi_t$

Define:

$$H_t = \sigma(\Delta\Psi_t / (N_t + \varepsilon))$$

where σ clips to $[0,1]$.

High H_t indicates aligned novelty; low H_t indicates brittle or disruptive change.

A.4 Phase Regions

Using these parameters, SFD phase regimes become measurable:

- **M-phase:** low Ψ and/or low A (fragmentation)
- **W-phase:** rising Ψ , sufficient A , high H (repair and coherence formation)
- **I-phase:** high stable Ψ , low N (persistent attractor)

Pathological coherence corresponds to high Ψ and A with declining H .

A.5 Scope and Intent

This appendix is not a claim of complete formalization. It is a minimal consistency proof: that the concepts introduced in the paper can be represented as concrete, measurable quantities without violating the framework's commitments.

Different domains may instantiate these quantities using different proxies; the formalization here illustrates structure, not prescription.

Further formal development, empirical testing, and domain-specific instantiation are left to future work.

A.6 Practical Instantiation of the Latent Space Z

A recurring question is how to interpret the latent semantic variable z in practical settings—particularly for domains as complex and contested as human values.

Within Semantic Field Dynamics, z is **not** assumed to be a universal or pre-defined semantic space. Instead, it is identified *operationally*, based on what structure coordination depends upon in a given interaction context.

General principle. z denotes whatever latent structure, when participants' constraints align within it, produces a measurable reduction in coordination cost. Its adequacy is determined by function, not ontology.

Put simply: z is *whatever space the interaction is implicitly organizing itself around*.

"The question of what z 'really is' is therefore less important than whether alignment within a proposed z consistently reduces coordination cost. Semantic Field Dynamics treats this functional criterion—not ontological commitment—as decisive."

Domain-specific examples

- **Human–AI dialogue.** In conversational settings, z may be instantiated as a semantic embedding space (e.g., sentence-level embeddings). Local constraint distributions $q_i(z)$ correspond to inferred intent, topic, or task interpretation. Field coherence corresponds to increasing embedding alignment across turns, which predicts fewer repair moves and greater interaction stability.
- **Multi-agent reinforcement learning.** In cooperative or mixed-motive environments, z may be a learned latent state or representation space internal to the agents. Alignment in this space corresponds to convergence on shared situational models or joint policies, reducing coordination failures and training instability.
- **Value alignment contexts.** For value-sensitive interaction, z may represent a preference or ordering space inferred through behavior (e.g., inverse reinforcement learning or choice modeling). Alignment in this space corresponds to improved prediction of choices and reduced corrective intervention, rather than agreement on explicit value labels.

Context dependence as a feature

The framework does **not** require pre-specifying a single, global semantic space. Different interactions may organize around different latent structures, and the same participants may coordinate over different z in different contexts.

This is not a limitation of the framework but a consequence of its relational commitments: semantic fields are context-sensitive, historically contingent, and task-dependent.

Validation criteria

A proposed instantiation of z is adequate if:

- Participants can be shown to maintain constraints over it (explicitly or implicitly),
- Increased alignment in z correlates with reduced coordination cost (e.g., fewer repairs, lower prediction error, greater stability),
- Measures of coherence defined over z improve predictive power for interaction outcomes.

If these conditions are met, z serves its functional role within the framework regardless of its representational form or ontological interpretation.

Locality and Plurality of Semantic Fields

An important implication of this operational definition is that alignment does not require a single, global latent space shared across all contexts or users. Instead, alignment emerges through many locally coherent semantic fields, each defined by the interaction context in which coordination occurs.

A human–AI dyad coordinating on medical decision-making, creative collaboration, or strategic planning will instantiate different latent spaces z , even when mediated by the same underlying model. This plurality is not a limitation but a scalability feature: coherence is achieved locally, where coordination costs are actually incurred, rather than enforced globally through centralized objectives.

From this perspective, large-scale alignment arises not from universal value encoding, but from the system's capacity to rapidly form, maintain, and revise many context-specific semantic fields with low coordination cost.

References

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.
- Bai, Y., Kadavath, S., Kundu, S., Askill, A., Kernion, J., Jones, A., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
- Buzsáki, G. (2020). The Brain from Inside Out. Oxford University Press.
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. arXiv:1706.03741.
- Epstein, I. R., & Pojman, J. A. (1998). An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos. Oxford University Press.
- Fries, P. (2015). Rhythms for Cognition: Communication through Coherence. *Neuron*, 88(1), 220–235.
- Hinton, G. E. (2002). Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8), 1771–1800.
- Hutchins, E. (1995). Cognition in the Wild. MIT Press.
- Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.
- Kessin, R. H. (2001). Dictyostelium: Evolution, Cell Biology, and the Development of Multicellularity. Cambridge University Press.
- Krakovna, V., et al. (2020). *Specification gaming: The flip side of AI ingenuity*. DeepMind Safety Research. <https://deepmind.com/research/publications/specification-gaming>
- Leibo, J. Z., et al. (2019). *Autocurricula and the emergence of innovation from social interaction*. *Nature Machine Intelligence*, 1, 333–343.
- Michod, R. E., & Roze, D. (2001). *Cooperation and conflict in the evolution of multicellularity*. *Heredity*, 86, 1–7.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv:2203.02155.
- Prigogine, I., & Stengers, I. (1984). Order Out of Chaos: Man's New Dialogue with Nature. Bantam Books.
- Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.
- Turing, A. M. (1952). The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641), 37–72.
- Varela, F., Lachaux, J.-P., Rodriguez, E., & Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2, 229–239.
- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004), 686–688.

Semantic Field Dynamics · Michael Kontorovich, December 2025. Published by Polyworld at polyworld.life/thinking/semantic-field-dynamics/