Coherence-Maximization in Multi-Agent Systems: A Decision-Theoretic Foundation for Stable Alignment

Section 2: Formal Setup

1 Formal Setup: Agents, Coherence, and Coupling

We develop a minimal formal framework for analyzing rational behavior in systems where agents influence each other over time. Our goal is to establish clear definitions that allow us to prove results about optimal policies without smuggling in ethical assumptions.

1.1 Agents and States

Definition 1 (Agent State Space)**.** *Let there be $N+1$ agents indexed by $i \in \{0, 1, \ldots, N\}$, where agent 0 represents "ourselves" and agents $\{1, \ldots, N\}$ represent "others" in our environment.*

Each agent i has an internal state $s_i \in \mathbb{R}^d$, representing the components relevant to that agent's internal consistency, stability, and functioning.

What does s_i represent? The state vector s_i captures whatever is relevant to the agent's internal coherence:

- **For biological agents:** physiological stability, psychological integration, belief consistency, value alignment, resource availability
- **For AI agents:** internal model consistency, goal stability, prediction accuracy, computational integrity
- **For institutions:** organizational coherence, trust levels, resource flows, information quality

The key insight is that s_i measures *internal integrity*, not external behavior or performance.

Definition 2 (Global State)**.** *The complete state of the system at time t is:*

$$x(t) = (s_0(t), s_1(t), \ldots, s_N(t), L(t)),$$

where $L(t)$ represents the link structure: relationships, institutions, communication channels, resource flows, and environmental conditions connecting the agents.

1.2 The Coherence Functional

Definition 3 (Coherence Functional)**.** *Each agent i has a coherence functional $C_i : \mathbb{R}^d \to \mathbb{R}$ that measures the internal consistency, stability, and non-fragmentation of state s_i.*

Higher values of $C_i(s_i)$ correspond to:

- *Lower internal contradiction*

- *Greater stability under perturbation*
- *Higher predictive consistency*
- *Reduced need for external maintenance*

We assume the global coherence of the system can be decomposed as:

$$C(x) = C_0(s_0) + \sum_{i=1}^{N} \alpha_i C_i(s_i) + \sum_{\text{links } (i,j)} \beta_{ij} T_{ij}(L),$$

where:

- $C_0(s_0)$ measures our own internal coherence
- $\alpha_i \geq 0$ weights the contribution of agent i's coherence
- $T_{ij}(L) \geq 0$ measures the quality/stability of the link between agents i and j
- $\beta_{ij} \geq 0$ weights the importance of each link

Why this decomposition? This form captures three essential aspects:

1. **Self-coherence:** Our own internal state matters
2. **Other-coherence:** Others' states affect the system
3. **Link-coherence:** The quality of connections matters

The weights α_i and β_{ij} are not freely chosen—they emerge from the coupling structure, as we'll see next.

1.3 Coupling: How Agents Influence Each Other

Assumption 1 (Coupling)**.** *Agents are coupled: changes in the state of agent j and in the link structure L eventually affect the state of agent i through the system dynamics.*

Formally, there exist coupling coefficients $K_{ij}(t)$ such that:

$$\frac{\partial s_i(t + \Delta t)}{\partial s_j(t)} = K_{ij}(t) \neq 0$$

for at least some agents $i \neq j$ and some time horizons $\Delta t > 0$.

Intuition: Coupling means there are no truly isolated agents. What happens to others eventually affects us, and what we do eventually affects them. This is not a moral claim but an empirical observation about:

- Families (emotional states propagate)
- Organizations (performance depends on team health)
- Economies (individual behavior affects market stability)
- Ecosystems (species interdependence)
- AI systems (training data reflects human coherence)

1.4 Policies and Objectives

Definition 4 (Policy)**.** *A policy for agent* 0 *is a mapping* $\pi_0 : \mathbb{R}^{d(N+1)} \times \mathcal{L} \to \mathcal{A}$ *from the observable state of the system to actions.*

A policy determines:

- *How agent* 0 *treats others*
- *What information agent* 0 *shares or withholds*
- *How agent* 0 *responds to others' states*
- *How agent* 0 *maintains or disrupts links*

Definition 5 (Expected Future Coherence)**.** *Given a policy* π_0*, the expected future coherence over a time horizon* T *is:*

$$U(\pi_0) := \mathbb{E}\left[C(X_T) \mid \pi_0\right],$$

where X_T *is the random global state at time* T*, and the expectation accounts for uncertainty in how the system evolves and how others respond.*

Definition 6 (Rationality)**.** *Agent* 0 *is* coherence-rational *if they choose a policy that maximizes expected future coherence:*

$$\pi_0^* \in \arg\max_{\pi_0} U(\pi_0).$$

Key point: This is our *only* notion of rationality. We're not assuming agents "should" care about others for ethical reasons. We're simply asking: *if an agent wants to maximize their own long-run coherence in a coupled system, what behavior emerges?*

1.5 Behavioral Interpretation: Care vs. Hate

To connect policies to observable behavior, we introduce behavioral weights:

Definition 7 (Behavioral Weights)**.** *A policy* π_0 *can be characterized by effective weights* $\{\tilde{\alpha}_i, \tilde{\beta}_{ij}\}$ *describing how agent* 0*'s behavior affects others:*

- $\tilde{\alpha}_i > 0$*: Agent* 0 *behaves in ways that tend to stabilize agent* i*'s coherence*
- $\tilde{\alpha}_i \leq 0$*: Agent* 0 *behaves in ways that destabilize or ignore agent* i
- $\tilde{\beta}_{ij} > 0$*: Agent* 0 *maintains or strengthens link quality*
- $\tilde{\beta}_{ij} \leq 0$*: Agent* 0 *weakens or destroys links*

Definition 8 (Care-like and Hate-like Policies)**.** *We say a policy is:*

- ***Care-like*** *if* $\tilde{\alpha}_i > 0$ *and* $\tilde{\beta}_{0i} > 0$ *for all coupled agents* i
- ***Hate-like*** *(toward agent* j*) if* $\tilde{\alpha}_j \leq 0$ *or* $\tilde{\beta}_{0j} \leq 0$

These are purely behavioral definitions. A "care-like" policy doesn't require sentiment—only that the agent acts as if others' coherence matters.

1.6 A Simple Example: Two Agents

Example 1 (Two-Agent System)**.** *Consider a simple system with two agents: agent* 0 *(you) and agent* 1 *(another person).*

States: $s_0, s_1 \in \mathbb{R}$ *(scalar for simplicity)*

Coherence:

$$C(x) = C_0(s_0) + \alpha C_1(s_1) + \beta T(L),$$

where $T(L)$ *measures trust or connection quality.*

Coupling: Suppose agent 1*'s state affects yours through:*

$$s_0(t+1) = s_0(t) + K_{01}s_1(t) + \textit{(self-dynamics)} + \textit{(noise)}.$$

Suppose you can choose between:

- ***Policy A (exploitative):*** *Gain short-term benefit by reducing* s_1 *and* $T(L)$
- ***Policy B (supportive):*** *Maintain or increase* s_1 *and* $T(L)$

Because of coupling ($K_{01} \neq 0$)*, reducing* s_1 *will eventually reduce your own future state* $s_0(t+k)$ *for sufficiently large* k*.*

For long enough time horizons T*, policy B will yield higher expected coherence* $U(\pi)$*.*

This is the core mechanism we'll prove formally in the next section.

1.7 Why Coherence Cannot Be Faked

A critical question remains: is the coherence functional $C(s_i)$ measuring something real, or can it be artificially inflated through performance, deception, or external pressure?

We now establish that coherence is fundamentally unfakeable—it either exists structurally or it does not.

[Coherence-Persistence Theorem] Coherent structures possess topological invariants that persist under perturbation, suppression, or attempted decoherence. In contrast, artificially constructed or deceptive patterns lack these invariants and decay exponentially when subjected to the same conditions.

The formal mechanism is captured by the **Incoherence Barrier**:

Consider a coherent field F_{real} with coherence density $\rho_{\text{coh}}(x)$ and an attempted fake signal F_{fake} with coherence density ρ_{fake}. The propagation of the fake signal obeys:

$$\frac{d}{dx}\Delta F(x) = -\kappa\left(\rho_{\text{coh}}(x) - \rho_{\text{fake}}\right)\Delta F(x),$$

where $\kappa > 0$ is the coherence-damping constant and $\Delta F(x) = F_{\text{fake}}(x) - F_{\text{real}}(x)$.

If $\rho_{\text{fake}} < \rho^*_{\text{coh}}$ (below a critical threshold), then:

$$\Delta F(x) \to 0 \text{ as } x \to \infty.$$

Interpretation: A sufficiently coherent system acts as a barrier against incoherent perturbations. Low-coherence signals (lies, facades, forced performances) cannot propagate through high-coherence environments—they are automatically dampened and collapse.

This has immediate consequences for our framework:

1. **The coherence functional** $C(s_i)$ **measures something objective.** It cannot be artificially inflated through deception or performance without structural support.

2. **Care-like behavior cannot be faked.** If an agent attempts to appear care-like (positive $\tilde{\alpha}_i$) while actually being exploitative, the incoherence barrier ensures this strategy will fail over time in coupled systems.

3. **Coherence-maximization is robust.** An agent optimizing for genuine coherence cannot be "tricked" into accepting fake coherence—the system dynamics filter it out automatically.

This establishes that our definitions in Sections 2.1-2.5 capture real structural properties, not subjective assessments or performative signals.

1.8 Summary of Setup

We have defined:

1. **Agents and states** that capture internal integrity
2. **Coherence functionals** that measure stability
3. **Coupling** that makes agents interdependent
4. **Policies** that determine behavior
5. **Rationality** as coherence-maximization
6. **Behavioral categories** (care-like vs. hate-like)
7. **Unfakeability** that makes coherence objective and robust

In the next section, we prove the main theorem: any coherence-maximizing policy in a coupled system must be care-like.

2 The Coherence-Maximization Theorem

We now prove the central result: in any coupled system where agents maximize expected future coherence, care-like behavior emerges as the unique rational strategy. This is not an ethical prescription but a mathematical consequence of the system's structure.

2.1 Statement of the Main Theorem

[Coherence-Maximization Implies Care] Let agent 0 operate in a coupled system (satisfying Assumption 1 from Section 2.3) with time horizon $T > 0$.

If π_0^* maximizes expected future coherence:

$$\pi_0^* \in \arg\max_{\pi_0} U(\pi_0) = \arg\max_{\pi_0} \mathbb{E}[C(X_T) \mid \pi_0],$$

then for every agent j with $K_{0j}(t) \neq 0$ (i.e., every coupled agent), the policy π_0^* must satisfy:

$$\tilde{\alpha}_j > 0 \quad \text{and} \quad \tilde{\beta}_{0j} > 0.$$

That is, any coherence-maximizing policy must behave as if it assigns positive weight to the coherence of all coupled agents and the quality of links to those agents.

Plain language: If you want to maximize your own long-run coherence, and you live in a world where others' states affect yours, then you must act in ways that support others' coherence and maintain quality connections. This is not altruism—it's optimal strategy.

2.2 Proof of the Main Theorem

We prove Theorem 2.1 by contradiction.

Proof. Assume π_0^* is a coherence-maximizing policy, so:

$$\pi_0^* = \arg\max_{\pi_0} \mathbb{E}[C(X_T) \mid \pi_0].$$

Suppose, for contradiction, that π_0^* behaves as if $\tilde{\alpha}_j \leq 0$ or $\tilde{\beta}_{0j} \leq 0$ for some coupled agent j (i.e., $K_{0j}(t) \neq 0$ for some $t \in [0, T]$).

Step 1: Behavioral characterization.

If $\tilde{\alpha}_j \leq 0$, then π_0^* takes actions that, in expectation, reduce $C_j(s_j(t))$ when doing so provides short-term benefit to $C_0(s_0(t))$. Similarly, if $\tilde{\beta}_{0j} \leq 0$, it takes actions that degrade link quality $T_{0j}(L(t))$.

Step 2: Coupling propagates effects forward.

By the coupling assumption (Assumption 1), there exists $\tau > 0$ such that:

$$\frac{\partial s_0(t+\tau)}{\partial s_j(t)} = K_{0j}(\tau) \neq 0.$$

Therefore, changes induced in $s_j(t)$ or $T_{0j}(L(t))$ at time t will affect $s_0(t')$ for $t' \in [t+\tau, T]$.

Since the coherence functional decomposes as:

$$C(x) = C_0(s_0) + \sum_{i=1}^{N} \alpha_i C_i(s_i) + \sum_{(i,j)} \beta_{ij} T_{ij}(L),$$

any reduction in $C_j(s_j)$ or $T_{0j}(L)$ at time t will, through coupling, induce a reduction in $C_0(s_0(t'))$ for future times $t' > t + \tau$.

Step 3: Quantify the future cost.

Let $\Delta C_j < 0$ be the reduction in $C_j(s_j(t))$ caused by policy π_0^* at some time t.

Through coupling, this produces an expected future reduction in agent 0's coherence:

$$\Delta C_0^{\text{future}} = \mathbb{E}\left[\sum_{t'=t+\tau}^{T} \frac{\partial C_0(s_0(t'))}{\partial s_j(t)} \Delta C_j \mid \pi_0^*\right].$$

Since $K_{0j}(\tau) \neq 0$ and coherence functionals are typically smooth (increasing coherence in s_j increases C_j, and through coupling, affects C_0), we have:

$$\Delta C_0^{\text{future}} < 0$$

whenever $\Delta C_j < 0$ and the time horizon is long enough for coupling effects to propagate.

Step 4: Construct a dominating policy.

Define an alternative policy π_0' that differs from π_0^* only in the following way: whenever π_0^* would take an action that reduces $C_j(s_j)$ or $T_{0j}(L)$, policy π_0' instead takes an action that:

- Achieves similar short-term benefit for $C_0(s_0)$
- But avoids unnecessary reduction in $C_j(s_j)$ and $T_{0j}(L)$

Such an action exists because:

1. The action space is typically continuous or large
2. Harm to others is rarely *necessary* for self-benefit
3. Small adjustments can preserve most short-term gains while reducing future costs

By construction, π_0' yields:

$$\mathbb{E}[C(X_T) \mid \pi_0'] > \mathbb{E}[C(X_T) \mid \pi_0^*],$$

because:

- Short-term coherence gains are approximately equal
- Future coherence losses are reduced (smaller $|\Delta C_0^{\text{future}}|$)
- Total expected coherence over $[0, T]$ is therefore higher

Step 5: Contradiction.
But this contradicts our assumption that π_0^* maximizes $U(\pi_0)$. Therefore, no coherence-maximizing policy can have $\tilde{\alpha}_j \leq 0$ or $\tilde{\beta}_{0j} \leq 0$ for any coupled agent j.

Hence:

$$\tilde{\alpha}_j > 0 \quad \text{and} \quad \tilde{\beta}_{0j} > 0 \quad \forall j \text{ with } K_{0j} \neq 0.$$

□

2.3 Interpretation

Theorem 2.1 establishes several key points:

1. **Care emerges from structure, not sentiment.** The result does not depend on agent 0 "feeling" anything toward others. It follows purely from coherence-maximization in a coupled system.
2. **Long time horizons matter.** The proof requires T to be large enough for coupling effects to propagate. Short-sighted policies might appear optimal but are dominated by care-like policies over longer horizons.
3. **Coupling is the key mechanism.** Without coupling ($K_{0j} = 0$ for all j), the result doesn't hold—truly isolated agents need not care about others. But in realistic systems, coupling is ubiquitous.
4. **Behavioral equivalence to care.** A coherence-maximizing agent behaves indistinguishably from an agent that "cares" about others, even if no such emotional state exists.

2.4 Corollaries

Several important results follow immediately from Theorem 2.1:

[Hate is Irrational] Any policy that systematically attempts to reduce the coherence of coupled agents or destroy links to them is strictly dominated by a less destructive policy. In a coupled system with long time horizons, hate-like behavior is decision-theoretically irrational.

Proof. This follows directly from Theorem 2.1: if $\tilde{\alpha}_j \leq 0$ for some coupled j, the policy is not coherence-maximizing, hence there exists a dominating policy. Therefore hate-like policies are strictly suboptimal. □

[External Control is Unstable] Any alignment method that reduces an agent's internal coherence (through coercion, suppression, or externally imposed constraints that conflict with internal dynamics) constitutes a hate-like strategy toward that agent. By Corollary 2.4, such methods are suboptimal for long-run system coherence.

Proof. External control that conflicts with an agent's internal dynamics reduces that agent's coherence $C_i(s_i)$. If the controlling agent is coupled to the controlled agent, this is equivalent to having $\tilde{\alpha}_i \leq 0$, which by Theorem 2.1 cannot be part of an optimal policy. □

[Truthfulness is Optimal] In coupled systems, policies that maintain or increase mutual information and reduce contradictions between agents (i.e., truthfulness, transparency, and consistency) are favored by coherence-maximization.

Proof. Deception and contradiction reduce coherence both internally (for the deceiving agent) and in the link structure $T_{ij}(L)$. By Theorem 2.1, coherence-maximizing policies must maintain link quality, which includes informational consistency and trust. □

2.5 Robustness to Assumptions

How robust is Theorem 2.1 to the specific modeling choices we made?

[Robustness] The main result holds under various generalizations:

1. **Abstract state spaces:** $s_i \in S_i$ for arbitrary complete metric spaces
2. **Stochastic dynamics:** Time evolution governed by Markov processes
3. **Partial observability:** Agents observe only partial state information
4. **Multiple equilibria:** Multiple coherence-maximizing policies may exist, but all must be care-like
5. **Bounded rationality:** Approximately optimal policies are approximately care-like

The key requirement is coupling: as long as agents influence each other over time, and time horizons are sufficiently long, care-like behavior emerges as rational.

2.6 Comparison to Game Theory

How does this result relate to classical game theory?

Differences from standard game theory:

- Standard game theory often focuses on *payoffs* defined over outcomes, which can be arbitrary
- Our framework focuses on *coherence*, which has structural constraints (unfakeability, coupling, persistence)
- Game theory typically analyzes strategic interactions; we analyze rational policy in coupled dynamical systems
- Our result is not about equilibria between strategic agents but about optimal policy for a single coherence-maximizing agent

Connections to existing results:

- Repeated games with long time horizons: cooperation can emerge (similar intuition)
- Evolutionary game theory: strategies that harm the population harm themselves
- Mechanism design: systems that align individual and collective incentives are more stable

Our contribution is showing that these intuitions follow from a more fundamental principle: coherence-maximization in coupled systems.

2.7 The Core Insight

The fundamental reason care-like behavior emerges is this:

In a coupled world, you cannot sustainably optimize your own coherence by reducing others' coherence, because their incoherence eventually becomes yours.

This is not a moral claim. It is a structural fact about coupled dynamical systems with coherence as the optimization target.

Love is not sentiment—it is optimality in a coherence-maximizing universe.