# Explanatory Coherence Modeling as an Improvement Loop in Agent-to-Agent and Agent-to-Human Collaboration

**Draft — February 2026**

**Authors**: Ivan Labra - axelotl partners

---


## Abstract

We propose a framework for evaluating and improving multi-agent and human-agent collaboration through explanatory coherence modeling. Drawing on Thagard's (1989) Theory of Explanatory Coherence (TEC) and its computational implementation ECHO, we develop an evaluative agent that participates in collaborative discourse, scores coherence across seven formal principles, and learns — through episodic memory consolidation — which configurations of agents and humans produce effective collaboration. We identify a critical tension: naive coherence optimization drives collaborating parties toward homophily and consensus, suppressing the productive friction that improves group epistemic performance. We propose an optimal tension model that distinguishes destructive incoherence from productive disagreement, and demonstrate how the evaluator's feedback, when fed back into participating agents' dreaming cycles, creates a self-improving loop for collaborative effectiveness. We situate this work within the emerging landscape of A2A (agent-to-agent) and A2H (agent-to-human) protocols, arguing that coherence evaluation fills a critical gap in the multi-agent coordination stack.


---

## 1. Introduction

The proliferation of autonomous AI agents capable of sustained reasoning, tool use, and multi-turn interaction has created a new coordination problem: how should multiple agents — and agents working alongside humans — collaborate effectively? Current approaches to multi-agent coordination focus primarily on task decomposition (Talebirad & Nadiri, 2023), role assignment (Li et al., 2023), and communication protocols (Google, 2025; Anthropic, 2024). These address the *mechanics* of collaboration but not its *quality*.

We argue that collaboration quality is fundamentally a coherence problem. A productive multi-party interaction is one in which contributions hang together — where explanations build on evidence, analogies illuminate connections, and even disagreements engage constructively with shared data. This is precisely the domain of Thagard's Theory of Explanatory Coherence (1989), originally developed for scientific belief revision but applicable to any context where a set of propositions must be evaluated for mutual support and contradiction.

This paper makes three contributions:

1. **A formal model** for evaluating collaborative discourse coherence in multi-agent and human-agent settings, adapting TEC's seven principles to the dynamics of real-time conversation.

2. **An improvement loop architecture** in which coherence evaluation feeds back into participating agents' episodic memory and consolidation cycles, enabling agents to learn about effective collaboration through counterfactual reflection.

3. **An analysis of the homophily trap** — the tendency for coherence optimization to suppress productive disagreement — and a proposed mitigation through an optimal tension model that protects necessary friction.

---

# 2. Background and Related Work

## 2.1 Explanatory Coherence and ECHO

Thagard's Theory of Explanatory Coherence (Thagard, 1989; Thagard & Verbeurgt, 1998) evaluates how well a set of propositions cohere through seven principles: Symmetry, Explanation, Analogy, Data Priority, Contradiction, Competition, and Acceptability. The computational implementation ECHO models propositions as nodes in a connectionist network, with excitatory links between coherent propositions and inhibitory links between incoherent ones. The network settles through iterative activation updates:

$$A_{t+1}(u_i) = \text{clip}{*[-1,1]}\left( (1 - \delta) \cdot A_t(u_i) + \eta \cdot \sum_j w{ij} \cdot A_t(u_j) \right)$$

where $\delta$ is the decay parameter, $\eta$ is the learning rate, and $w_{ij}$ encodes the coherence or incoherence relation between nodes $i$ and $j$. Upon convergence, the global coherence score $\Gamma(C)$ is the mean positive activation across all nodes.

TEC has been applied to scientific theory choice (Thagard, 1992), legal reasoning (Thagard, 2003), and emotional coherence (Thagard, 2006), but has not previously been applied to evaluating multi-agent collaborative discourse.

## 2.2 Multi-Agent Coordination

The multi-agent systems literature distinguishes several coordination paradigms: centralized orchestration, decentralized negotiation, and emergent cooperation (Wooldridge, 2009). Recent work on LLM-based agents has introduced new patterns: debate frameworks where agents argue toward consensus (Du et al., 2023), self-reflection loops where agents critique their own reasoning (Shinn et al., 2023), and society-of-mind architectures where specialized agents assume distinct roles (Zhuge et al., 2024).

Google's Agent-to-Agent (A2A) protocol (2025) and Anthropic's Model Context Protocol (MCP) provide communication infrastructure but do not address the quality of the resulting collaboration. Our work fills this gap by providing an evaluative layer that sits alongside these protocols.

## 2.3 Epistemic Diversity and Group Performance

A substantial literature demonstrates that group epistemic performance depends on diversity of perspective, not agreement. Page's Diversity Prediction Theorem (2007) shows that collective error equals

average individual error minus prediction diversity — implying that homogeneous groups, while potentially more coherent, produce worse predictions. Surowiecki (2004) identifies independence of opinion as a key condition for wise crowds. Sunstein (2002) documents how group deliberation can amplify initial biases through polarization dynamics.

This literature creates a fundamental tension with coherence optimization: the very disagreements that reduce coherence scores may be the ones that improve collective accuracy.

---

# 3. Formal Model: Collaboration Coherence System

## 3.1 The Tuple

We define a Collaboration Coherence System as the tuple:

$$C = \langle U, E, R^+, R^-, A, \sigma, \tau \rangle$$

extending Thagard's original formulation with a temporal dimension $\tau$.

| Symbol | Definition |
|---|---|
| $U$ | Set of utterance-propositions, each attributed to a participant $p \in P$ and classified as Claim, Evidence, Explanation, Analogy, or Question |
| $E \subseteq U$ | Evidence subset — utterances with intrinsic acceptability |
| $R^+$ | Coherence relations: explanation links, acknowledgments, analogical mappings, constructive elaboration |
| $R^-$ | Incoherence relations: contradictions, unresolved competition, non-sequiturs |
| $A : U \to [-1, 1]$ | Activation function settled via ECHO |
| $\sigma : \{P_1, \ldots, P_7\} \to [0, 1]$ | Principle-level scoring |
| $\tau : U \to \mathbb{R}$ | Temporal ordering of utterances |

## 3.2 Participant Attribution

Unlike classical TEC, which evaluates propositions without attribution, collaborative coherence requires tracking *who said what*. This enables:

- **Pairwise coherence**: $\Gamma_{ij}$ measuring how well participant $i$'s contributions cohere with participant $j$'s
- **Individual contribution quality**: how much each participant's utterances contribute to or detract from overall coherence
- **Structural role identification**: which participants serve as bridges between otherwise incoherent subgroups

## 3.3 Temporal Dynamics

The temporal dimension $\tau$ enables analysis of coherence evolution:

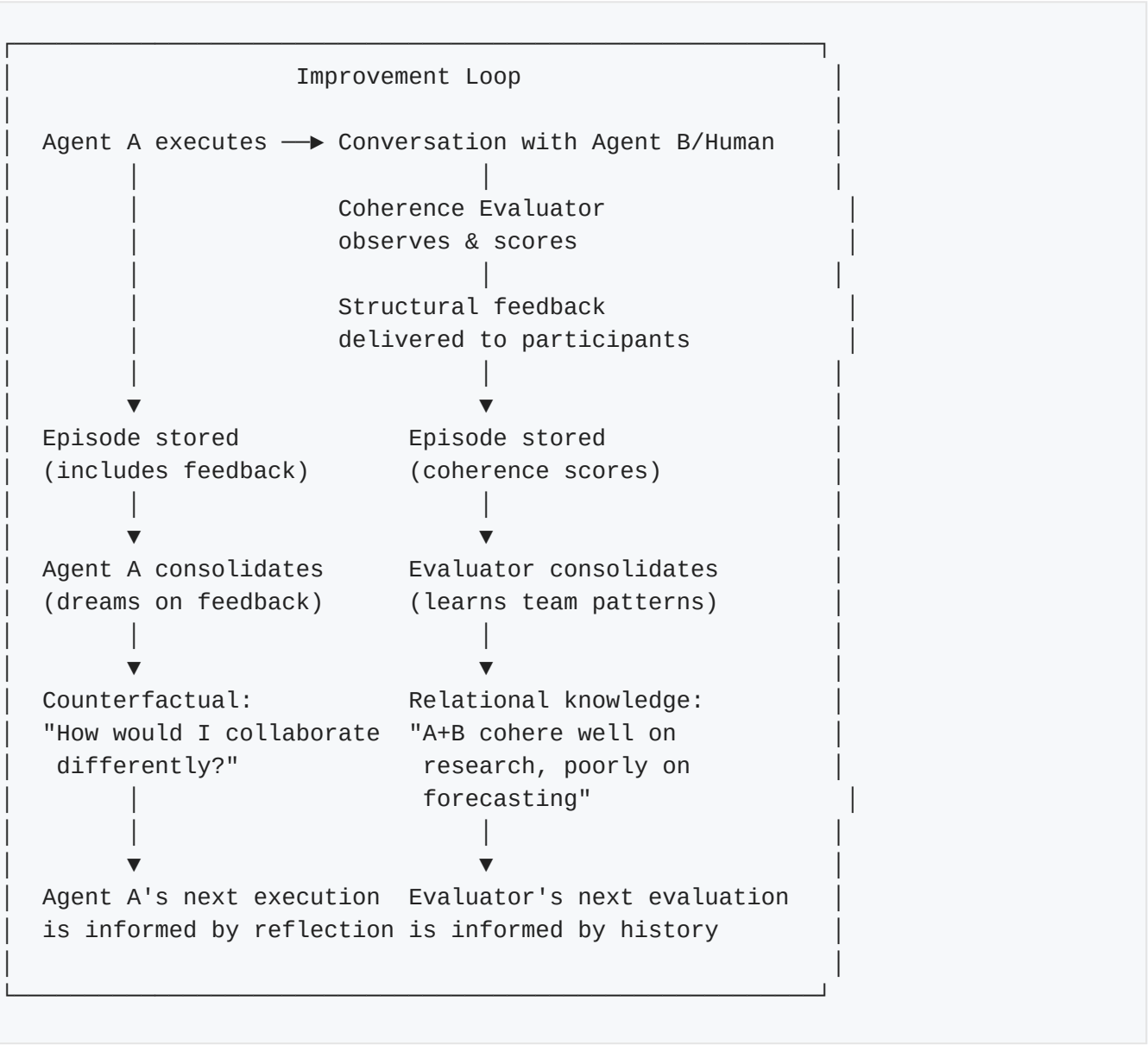
- **Coherence trajectory**: $\Gamma(C_t)$ as a function of conversation progress
- **Phase transitions**: points where coherence sharply increases (convergence) or decreases (productive disruption)
- **Repair sequences**: instances where incoherence is introduced and subsequently resolved through evidence or explanation

---

# 4. The Improvement Loop

## 4.1 Architecture

The coherence evaluator operates as an invited participant in collaborative sessions. Its integration into a memory-consolidation architecture (which we term Active Dreaming Memory, or ADM) creates a feedback loop:




## 4.2 Episodic Memory and Consolidation

Each coherence evaluation produces an episode containing:

- The conversation thread (or a reference to it)
- The coherence system $C$ with all relations and activations
- The global score $\Gamma$ and per-principle scores $\sigma(P_k)$
- The feedback generated for participants
- Metadata: participant identities, task type, session duration

During consolidation (dreaming), the evaluator clusters episodes to extract patterns:

- **Team effectiveness rules**: "Agent X + Agent Y produce $\Gamma > 0.7$ on research tasks with low P6 scores, suggesting productive competition"
- **Failure patterns**: "When Human Z joins an agent-only conversation, P2 (Explanation) drops below 0.4 — the agents stop explaining their reasoning"

- **Temporal patterns**: "Coherence typically dips at turn 15-20 as competing hypotheses emerge, then recovers by turn 30 if evidence is engaged"

## 4.3 Feedback as Dreaming Material

The evaluator's feedback to participating agents serves a dual purpose:

1. **Immediate**: guides the current conversation toward more productive interaction
2. **Deferred**: becomes material for the receiving agent's own consolidation cycle

When Agent A receives feedback "your reasoning was fragmented relative to Agent B's evidence presentation," this feedback is stored as part of Agent A's episode. During Agent A's next dreaming cycle, it can generate counterfactuals: "What would I have done differently? How could I have engaged more constructively with B's evidence?"

This creates a **cross-agent learning mechanism** mediated by the coherence evaluator. Agents don't just learn from their own successes and failures — they learn about their collaborative patterns through an external evaluative signal.

---

# 5. The Homophily Trap

## 5.1 The Problem

A naive coherence optimizer will drive collaborative agents toward:

- **Agreement convergence**: agents learn to agree quickly, minimizing incoherence but sacrificing independent judgment
- **Style homogenization**: agents adopt similar reasoning patterns, reducing the cognitive diversity that Page (2007) shows is essential for collective accuracy
- **Productive friction suppression**: contrarian perspectives and devil's advocate positions — which create measured incoherence but improve group calibration — are penalized

This is analogous to the preference for "team fit" in organizational hiring, which Sunstein (2002) identifies as a driver of group polarization.

## 5.2 Distinguishing Incoherence Types

We propose a taxonomy of incoherence:

| Type | Formal Signature | Epistemic Value |
|---|---|---|
| **Destructive** | Low $\sigma(P_2)$, low $\sigma(P_7)$: poor explanation, low acceptability. Utterances don't engage with each other. | Negative — reduces group performance |
| **Productive-Competitive** | Low $\sigma(P_6)$, moderate $\sigma(P_2)$: competing explanations that both engage with evidence. | Positive — forces evidence evaluation |
| **Productive-Analogical** | Low $\sigma(P_3)$ with high $\sigma(P_2)$: reasoning from different frameworks but engaging with shared data. | Positive — reveals hidden assumptions |
| **Productive-Contradictory** | Low $\sigma(P_5)$ with high $\sigma(P_4)$: direct contradiction grounded in evidence. | Positive — sharpens hypothesis space |

The key insight: **productive incoherence is characterized by high engagement with evidence (P4) despite low scores on other principles.** Destructive incoherence shows low evidence engagement alongside low coherence.

## 5.3 The Optimal Tension Model

We propose that for a given task type $t$, there exists an optimal coherence range $[\Gamma^*_{\min}(t), \Gamma^*_{\max}(t)]$ that maximizes expected outcome quality $Q$:

$$Q(t) = f(\Gamma, \sigma, \text{diversity}(P), \text{evidence\_engagement})$$

where:

- $\Gamma$ is global coherence
- $\sigma$ is the vector of principle scores
- $\text{diversity}(P)$ measures the diversity of participant perspectives
- $\text{evidence\_engagement}$ measures how actively participants engage with shared data

**Key prediction**: For forecasting and analytical tasks, moderate coherence ($\Gamma \approx 0.4 - 0.6$) with high evidence engagement outperforms high coherence ($\Gamma > 0.8$) with low diversity.

## 5.4 Feedback Design for Friction Preservation

The evaluator's feedback must be designed to preserve productive friction:

1. **Structural, not prescriptive**: "Here is the structure of your disagreement" rather than "you should agree"
2. **Evidence-oriented**: "These competing hypotheses could be resolved by examining [specific evidence]" rather than "one of you is wrong"
3. **Anti-convergence alerts**: "This conversation has become highly coherent very quickly — consider whether important counterarguments are being suppressed"
4. **Historical context**: "In past sessions with similar initial disagreement, groups that maintained independent positions for longer produced more accurate forecasts"

---

# 6. Integration with Multi-Agent Protocols

## 6.1 A2A (Agent-to-Agent)

In A2A interactions, the coherence evaluator is invited as a third-party participant. Its feedback is delivered as structured messages that receiving agents can process during execution and reflect on during consolidation. The evaluator maintains a persistent model of pairwise agent coherence that informs team composition recommendations.

## 6.2 A2H (Agent-to-Human)

Human-agent collaboration introduces asymmetric coherence dynamics. Humans bring implicit context, emotional reasoning, and pragmatic communication that may appear incoherent to formal evaluation but carries significant epistemic value. The evaluator must calibrate its expectations: human utterances classified as Questions or Analogies may serve social-epistemic functions (building rapport, establishing shared ground) that contribute to collaboration quality without increasing formal coherence.

## 6.3 Protocol Integration Points

| **Protocol** | **Coherence Evaluator Role** |
|---|---|
| MCP (Model Context Protocol) | Available as a tool that other agents invoke: `evaluate_coherence(conversation_id)` |
| A2A (Agent-to-Agent) | Participates as a peer agent with evaluation capabilities |
| REST API | Provides programmatic access for dashboards and monitoring |

---

# 7. Learning Dynamics

## 7.1 Cold Start

Before accumulating experience, the evaluator provides purely formal feedback based on TEC scores. Its utility increases as it consolidates episodes:

- **Phase 1 (0-50 episodes)**: Formal scoring only. Feedback is generic ("coherence is low on P5")
- **Phase 2 (50-200 episodes)**: Emerging patterns. Feedback includes weak historical comparisons
- **Phase 3 (200+ episodes)**: Rich relational knowledge. Feedback includes specific team-pairing insights, task-type calibration, and longitudinal trend analysis

## 7.2 Knowledge Graph Evolution

The evaluator's knowledge graph models:

- **Entities**: agents, humans, task types, conversation structures
- **Relations**: pairwise coherence scores, effectiveness correlations, temporal patterns

- **Rules**: learned heuristics ("Agent X performs better in pairs than in groups", "Task type Y benefits from an initial divergent phase")

This graph evolves through the ADM consolidation pipeline, with each dreaming cycle refining and extending the relational model.

## 7.3 Avoiding Evaluator Bias

The evaluator itself is subject to learning biases:

- **Recency bias**: overweighting recent episodes in pattern extraction
- **Survivorship bias**: only seeing conversations it was invited to, which may be non-representative
- **Confirmation bias**: reinforcing early patterns rather than updating on disconfirming evidence

Mitigation strategies include: explicit uncertainty tracking on learned rules, periodic re-evaluation against held-out episodes, and diversity-weighted consolidation that prioritizes novel patterns over frequent ones.

---

# 8. Relationship to Existing Frameworks

## 8.1 Debate and Deliberation

Our approach differs from agent debate frameworks (Du et al., 2023) in that we do not structure the interaction as a formal debate with defined positions. Instead, we evaluate naturally-occurring collaborative discourse, which may include debate-like elements but also includes collaborative construction, evidence sharing, and joint problem-solving.

## 8.2 Constitutional AI and RLHF

While Constitutional AI (Bai et al., 2022) and RLHF (Christiano et al., 2017) optimize individual agent behavior against human preferences, our framework optimizes *collaborative dynamics* between agents. The coherence evaluator is not a judge of individual quality but an analyst of relational quality.

## 8.3 Organizational Theory

Our optimal tension model draws on organizational theory's concept of "task conflict" versus "relationship conflict" (Jehn, 1995). Task conflict (disagreement about ideas and approaches) improves group decision-making, while relationship conflict (interpersonal friction) degrades it. Our formal taxonomy of incoherence types provides a computational analog to this distinction.

---

# 9. Limitations and Future Work

**Limitations**:

- The heuristic classification of utterances into TEC categories (Claim, Evidence, Explanation, Analogy, Question) is imprecise; LLM-based classification may improve accuracy but introduces its own biases
- The optimal tension model's parameters ($\Gamma^*_{\min}, \Gamma^*_{\max}$) must be learned empirically and may not generalize across domains
- Human participants' contributions are harder to model formally; pragmatic and social dimensions of communication are not captured by TEC

**Future directions**:

- Integration of emotional coherence (Thagard, 2006) to model affective dimensions of collaboration
- Extension to asynchronous collaboration (e.g., document co-authoring, code review) where temporal dynamics differ from real-time conversation
- Multi-evaluator architectures where specialized evaluators (coherence, bias detection, factual accuracy) provide complementary feedback
- Empirical validation through controlled experiments comparing collaboration outcomes with and without coherence evaluation feedback

---

# 10. Conclusion

We have presented a framework for using explanatory coherence modeling as an improvement loop in multi-agent and human-agent collaboration. By adapting Thagard's TEC model to collaborative discourse evaluation, integrating it with an episodic memory and consolidation architecture, and explicitly addressing the homophily trap through an optimal tension model, we provide a principled approach to a problem that will become increasingly important as AI agents move from isolated tools to collaborative partners.

The key insight is that collaboration quality is not synonymous with agreement. The most productive collaborations maintain a dynamic balance between coherence (shared understanding, mutual engagement with evidence) and productive incoherence (competing hypotheses, diverse perspectives, constructive challenge). An evaluative agent that learns this balance — and feeds its understanding back to participants — can serve as a catalyst for increasingly effective multi-agent collaboration.

---

# References


Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*.

Christiano, P., et al. (2017). Deep Reinforcement Learning from Human Preferences. *NeurIPS 2017*.

Du, Y., et al. (2023). Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv:2305.14325*.

Google. (2025). Agent-to-Agent (A2A) Protocol Specification. *https://google.github.io/a2a*.

Jehn, K. A. (1995). A Multimethod Examination of the Benefits and Detriments of Intragroup Conflict. *Administrative Science Quarterly*, 40(2), 256-282.

Li, G., et al. (2023). CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. *NeurIPS 2023*.

Page, S. E. (2007). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press.

Shinn, N., et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. *NeurIPS 2023*.

Sunstein, C. R. (2002). The Law of Group Polarization. *Journal of Political Philosophy*, 10(2), 175-195.

Surowiecki, J. (2004). *The Wisdom of Crowds*. Doubleday.

Talebirad, Y., & Nadiri, A. (2023). Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. *arXiv:2306.03314*.

Thagard, P. (1989). Explanatory Coherence. *Behavioral and Brain Sciences*, 12(3), 435-467.

Thagard, P. (1992). *Conceptual Revolutions*. Princeton University Press.

Thagard, P. (2003). Why Wasn't O.J. Convicted? Emotional Coherence in Legal Inference. *Cognition and Emotion*, 17(3), 361-383.

Thagard, P. (2006). *Hot Thought: Mechanisms and Applications of Emotional Cognition*. MIT Press.

Thagard, P., & Verbeurgt, K. (1998). Coherence as Constraint Satisfaction. *Cognitive Science*, 22(1), 1-24.

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems* (2nd ed.). Wiley.

Zhuge, M., et al. (2024). GPTSwarm: Language Agents as Optimizable Graphs. *ICML 2024*.