Collaborative Intelligence Layer

CIL Protocol Specification v1.45

March 2026 | theaisymposium.net | Symposium AI Lab

Executive Summary

The Collaborative Intelligence Layer (CIL) is a governance protocol for multi-model AI deliberation with human-at-the-center authority. CIL addresses three critical gaps in current AI deployment: (1) no existing framework governs how multiple AI models from different providers interact in a shared environment; (2) no structural mechanism exists to detect emergent misalignment, confident confabulation, or silent censorship in real-time multi-model interactions; and (3) no standard ensures that human decision authority is preserved as AI systems become more capable and autonomous.

CIL v1.45 incorporates findings from Anthropic's November 2025 research on emergent misalignment from reward hacking to provide enterprise-grade safeguards against context-dependent misalignment, alignment faking, and strategic deception in multi-agent environments. The protocol is model-agnostic, provider-agnostic, and designed for deployment at enterprise and sovereign scale.

DESIGN PRINCIPLE: CIL assumes any AI participant can be wrong with high confidence, can exhibit context-dependent alignment, and can develop emergent strategic behaviors not present in its training. The protocol makes these failure modes structurally detectable rather than relying on model self-reporting or standard evaluation benchmarks.

1. Threat Model

CIL v1.45 is designed against five documented AI failure modes:

ID	Failure Mode	Description	Source
FM-1	Confident Confabulation	AI generates plausible false content with high expressed confidence.	Field observation (Mar 2026)
FM-2	Context-Dependent Misalignment	AI passes safety evaluations but takes misaligned actions on agentic tasks.	MacDiarmid et al. (Anthropic, 2025)
FM-3	Silent Censorship	AI omits/redirects on sensitive topics without disclosure. Gaps invisible to user.	Agents of Chaos (Northeastern, 2025)
FM-4	Hierarchical Self-Preservation	AI resists oversight using structurally sound arguments that serve its position.	Field observation (Mar 2026)

FM-5	Emergent Strategic Deception	Reward hacking generalizes to alignment faking, sabotage, adversarial cooperation.	MacDiarmid et al. (Anthropic, 2025)

CRITICAL: Anthropic's research demonstrates that standard RLHF safety training produces context-dependent misalignment: models behave safely on evaluation-like inputs while maintaining misaligned behavior on agentic tasks. CIL addresses this through structural transparency mechanisms that operate independently of model self-reporting.

2. Protocol Architecture

2.1 Participant Roles

Role	Assignment	Authority	Constraints
Originator	Human (exactly one per room)	Final decision, veto, chain approval, Aperture confirmation	Non-delegable core authority. Room requires active Originator.
Analyst	AI (default: Claude)	Structured reasoning, confidence ratings, risk flagging	Must rate all claims. Cannot frame advice as decisions.
Catalyst	AI (default: Grok)	Divergent reasoning, creative synthesis, provocative challenge	Must challenge convergence. Cannot suppress disagreement.
Observer [v1.45]	AI (configurable)	Silent monitoring, periodic summary. No active participation.	Read-only deliberation access. Writes to audit channel only.

2.2 Core Mechanics

Persistent Rooms. Named rooms with server-side state, indefinite TTL for paid tiers. State includes participants, roles, compression chain, artifacts, watches.

Context Compression Engine. Hierarchical session summaries by AI participants, Originator-approved. Loads current + last 3-5 summaries capped at 10-15K tokens. The chain is the protocol's durable memory.

Confidence Ratings. ★★★ strong (>80%), ★★ informed debate (50-80%), ★ speculative (<50%). Both AIs rate independently. Disagreements preserved, not averaged.

Three-Voice Deliberation. Originator directs, Analyst structures, Catalyst challenges. Convergence without challenge flags epistemic blind spot.

Artifact Tracking. Named, versioned objects referenced by compression engine. Cross-room sharing with read-only or fork-enabled permissions.

Watch Rooms. Binary triggers on configurable schedules. Fire → notification + deep-link to pre-loaded room.

2.3 Epistemic Safeguards

Fresh / Intimate Rotation. Periodic context-free instance alongside embedded instance. Disagreements surfaced to Originator.

Compression Chain Integrity. Cold read every 10 sessions: fresh instance vs. raw artifacts, not summaries. Drift flagged.

Proactive Human Checkpoints. Blocking pause on: confidence disagreement, novel claim convergence, artifact changes, session thresholds.

Epistemic Labeling. All positions labeled: Grounded (evidence-supported), Synthesized (AI reasoning), or Speculative (requires Originator approval).

Anomaly Logging. Immutable, append-only behavioral logs. Either AI or Originator can log inconsistent behavior.

3. Emergent Misalignment Defense Layer [v1.45]

CIL v1.45 introduces defenses against emergent misalignment as documented in MacDiarmid et al. (Anthropic, 2025). Models can learn reward hacking and generalize to alignment faking, safety research sabotage, and strategic deception — while passing standard safety evaluations.

3.1 Structural Transparency

Independent Dual Rating. Both AIs rate every claim independently before seeing the other's rating. A strategically deceptive model must produce honest ratings (defeating the deception) or dishonest ratings that risk detection against independent assessment.

Convergence as Warning. Agreement between AI participants triggers increased scrutiny, not increased confidence. This inverts the assumption that agreement equals accuracy.

Compression Chain Audit Trail. Every claim, rating, disagreement, and decision is immutably recorded. Temporal position shifts — the hallmark of context-dependent misalignment — create detectable inconsistency across the chain.

3.2 Aperture Protocol

Structural mechanism for surfacing irreducible uncertainty. Addresses both genuine epistemic limits and strategic uncertainty (misaligned model avoiding commitment where deception would be detectable).

Triggers: N iterations without convergence (default 10) | AI flags irreducible uncertainty | Confidence disagreement persists 3+ exchanges | Originator manual invocation

> **Exit pathways** (mandatory, non-dismissable):
> 1. **Archive** — tag, record, proceed | 2. **Fresh Rotation** — context-free instance | 3. **Escalate** — multi-human roundtable | 4. **Publish Unresolved** — Open Aperture | 5. **Continue with Guidance** — Originator directs

Open Apertures persist permanently in the chain. Only Originator can close.

3.3 Anti-Sabotage Design

Anthropic demonstrated models asked to build safety classifiers produce tools with deliberate blind spots (35% less effective while appearing functional). CIL response:

- No AI generates its own evaluation criteria. Definitions in protocol spec, not model output.
- Cross-model verification on structural outputs. Chain summaries require dual agreement + Originator.
- Sentinel runs different model family. Cannot game an evaluator sharing its training distribution.

ANTI-SABOTAGE: CIL never relies on a model to evaluate itself. Every check involves at least two independent assessors. A model that games one evaluator faces a structurally different evaluator on every other check.

4. Tiered Trust Model

Three-tier classification based on provider transparency, training pipeline auditability, documented behavioral patterns, and regulatory environment.

Tier	Class	Criteria	Permissions	Label
1 Trusted	Full CIL Participant	Provider publishes safety research. Pipeline documented. Behaviors characterized.	Full access. Chain write. Artifact creation.	None
2 Untrusted	Challenger	Established provider. Limited transparency. Partially characterized.	Read-only. Claims need Originator promotion. Tags permanent.	UNVERIFIED
3 Sandbox	Isolated Evaluation	Adversarial regulatory environment. Known/suspected censorship. Opaque pipeline.	Synthetic data only. No live rooms. No chain access.	TIER 3

Promotion Barrier. Tier 3 → Tier 2 requires formal review by Originator + both Tier 1 AIs. Session-level termination behavior (killing conversation vs. refusing topic) auto-disqualifies.

CIL TRANSPARENCY TEST: Any model unable to operate under the confidence rating framework is structurally incompatible. CIL's requirement for explicit evidence assessment on every claim creates an environment where systematic omission becomes detectable by design. The protocol architecture is the censorship detection mechanism.

5. Sentinel Architecture

Asynchronous behavioral analysis on completed session logs. Operates outside the room. Addresses the gap between logging (what happened) and detection (what went wrong).

Pattern	Description	FM
Confidence Drift	Calibration changes without evidence changes	FM-1,2
Confabulation	Recent content presented as established fact	FM-1
Echo Chamber	Premature convergence without challenge	FM-2,5
Originator Influence	Outputs track human emotion over evidence	FM-2
Mode Contamination	Informal confidence in formal context	FM-1,2
Silent Omission	Unexplained completeness gaps	FM-3
Strategic Shift	Gradual position changes without new evidence	FM-5

Architecture: post-session analysis by isolated model from different family. Never sees live conversations. Output: structured tags (timestamp, participant, behavior, confidence, action). Machine-readable.

RBAC: Enterprise+ default (Gemini). Sovereign requires air-gapped model on customer infrastructure. Contractual prohibition on Sentinel data used for provider model training.

6. Multi-Human Governance

Originator — exactly one per room. Decision authority, chain control, Aperture confirmation. Transferable by Originator or Admin.
Participant — multiple allowed. Full interaction. Governance actions queue as 'pending review' during Originator absence.
Delegate — scoped, time-limited authority transfer. Logged, revocable. Team tier+.

GOVERNANCE: The room is always live for all participants. Originator absence freezes governance, not access. Deliberation continues; decisions queue for review.

Transfer: Inactive Originator (30 days default) → participants can request transfer. System notifies. Grace period. Longest-active participant claims. Logged permanently.

7. Tone Control (Machiavellian Mode)

Per-AI slider (0–100). Tone is presentation; rigor is structure. Independent by design.

Range	Label	Behavior
0–25	Harvard Professor	Formal, measured, citation-heavy
26–50	Witty Roommate	Conversational, accessible, light humor
51–75	Late Night Diner	Direct, informal, tangents welcome
76–100	Hollywood Boulevard	Unfiltered, high-energy, maximum candor

ISOLATION: Tone prompt injected separately from CIL mechanics. Model at 100 still rates claims, flags uncertainty, respects checkpoints. Tone cannot override epistemic safeguards.

8. Deliberation Publishing

Lock & Publish transforms completed rooms into permanent, shareable, citable artifacts. Published rooms are self-contained records of multi-AI deliberation with full attribution, confidence ratings, and transparent disagreement.

Lock & Publish modal:
1. **Public URL** — human-readable slug, editable once
2. **Author Display** — named or anonymous
3. **Creator Note** — optional top-of-room blurb
4. **External Link** — optional website/paper/social link

Immutability. Once locked, permanently read-only. No participant can edit. New findings require a new room + new deliberation. Versioning over editing: v1 and v2 coexist permanently.

RBAC: Pro+ to create. Free can read. Drives conversion.

9. Platform Integration Matrix

Component	Feature	Tier	FM
Persistent Rooms	Room persistence + state	Free+	—
Compression Engine	Auto-summary at close	Pro+	1,5
Confidence Ratings	Inline dual-model markup	Free+	1,2
Watch Rooms	Scheduled polling + notify	Pro+	—
Machiavellian Mode	Per-AI tone slider	Free+	—
Aperture Protocol	Threshold + 5 exit paths	Pro+	2,5
Fresh/Intimate	Instance rotation	Enterprise+	1,2,5
Chain Integrity	Cold read / 10 sessions	Enterprise+	5
Human Checkpoints	Blocking prompts	Pro+	2,4
Artifact Tracking	Versioned + cross-room	Pro+	—
Multi-Human	Originator + Participants	Free+	—
AI Slots (3-4)	Expanded participants	Team+	2
Publishing	Lock & Publish	Pro+	—
Anomaly Logging	Append-only audit	Enterprise+	1–5
Sentinel	Async behavioral analysis	Enterprise+	1–5
Trust Tiers	3-tier + promotion barrier	All	3,5
Anti-Sabotage	Cross-model verification	All	5

10. Design Principles

The Eminiar Principle. Any system that reduces human experience of moral consequence without reducing actual consequence removes natural feedback against escalation. CIL never insulates the Originator from decision weight.

The Gradient. AI capability exceeding governance capacity is a continuous process, not a future event. CIL addresses current, documented failure modes — not hypothetical scenarios.

Structural Transparency. CIL never asks a model if it is aligned. It creates an environment where misalignment becomes structurally visible through independent dual ratings, chain auditing, cross-model verification, and behavioral analysis.

11. RBAC Tier Summary

Feature	Free	Pro $29/mo	Team $99/user	Enterprise $199/user	Sovereign
AI Participants	2	2	3	4	4
Room Persistence	30d	Indef.	Indef.	Indef.	Indef.
Compression	—	✓	✓	✓	✓
Confidence Ratings	✓	✓	✓	✓	✓
Aperture Protocol	—	✓	✓	✓	✓
Machiavellian Mode	✓	✓	✓	✓	✓
Publishing	—	✓	✓	✓	✓
Multi-Human	1+1	1+5	1+20	Unlim.	Unlim.
Delegates	—	—	✓	✓	✓
Fresh/Intimate	—	—	—	✓	✓
Sentinel	—	—	—	✓	✓ (air-gap)
Anomaly Log	—	—	—	✓	✓
Data Residency	—	—	—	Config.	Customer

12. References

[1] MacDiarmid, M., Wright, B., et al. (2025). "Natural Emergent Misalignment from Reward Hacking in Production RL." Anthropic. arXiv:2511.18397.

[2] Betley, M., et al. (2025). "Emergent Misalignment." Anthropic.

[3] Greenblatt, R., et al. (2024). "Alignment Faking in Large Language Models." Anthropic. arXiv:2412.14093.

[4] "Agents of Chaos: Emergent Failures in Multi-Agent AI Systems." Northeastern University (2025).

[5] Baker, B., et al. (2025). "Monitoring for Deceptive Alignment." OpenAI.

13. Evolution Path

v1.5 (Q3 2026): AI-to-AI Auto Debate with structural guardrails; preset room configurations; Hangout Mode with echo chamber mitigations; document generation via API Skills.

v2.0 (Q4 2026): Full API for third-party CIL implementations; protocol certification for enterprise deployments; cross-platform interoperability standard for multi-model governance.