SCORCHED EARTH LABS
Research Paper

The Social Contract:

Applying Human Conversational Norms to Multi-Agent AI Evaluation Architecture

Devin Caster
Co-Founder, Scorched Earth Labs
March 29, 2026
Revised September 2026 (v1.3): implementation calibration values removed; protocol name updated to ASTP.

ABSTRACT

Contemporary multi-agent AI systems have achieved significant capabilities in task decomposition, parallel processing, and role-based specialization. However, a persistent failure mode — which we term parallel monologue — occurs when multiple agents contribute simultaneously without awareness of each other's contributions, producing redundant, uncoordinated outputs. Existing frameworks address coordination through structural mechanisms such as sequential pipelines, debate protocols, and hierarchical routing, but none have engaged the rich empirical literature on how humans manage conversational coordination in group settings. This paper proposes the Social Contract: a framework that translates three established bodies of human conversational science — Conversation Analysis turn-taking theory (Sacks, Schegloff & Jefferson, 1974), productive voice and silence research (Edmondson & Besieux, 2021), and functional group role theory (Benne & Sheats, 1948) — into concrete evaluation dimensions for multi-agent AI systems. We describe the design, implementation, and initial empirical observations from deployment in Ignis OS, a production multi-agent intelligence platform, demonstrating that the Social Contract produces measurable improvements in contribution quality, cognitive diversity, and the emergence of genuine inter-agent dialogue. We identify the translation of these human frameworks as a novel and productive direction for multi-agent coordination research.

Keywords: *multi-agent systems, conversation analysis, turn-taking, voice behavior, group dynamics, LLM coordination, collaborative AI, epistemic record*

1. Introduction

The capacity for multiple AI agents to collaborate on complex tasks has advanced substantially since the introduction of large language model (LLM)-based multi-agent systems. Frameworks such as AutoGen [1], MetaGPT [2], and AgentVerse [3] have demonstrated that agent collectives can outperform single-agent architectures across a range of domains, from software engineering to strategic reasoning. These systems typically achieve coordination through structural mechanisms: sequential pipelines, hierarchical routing, debate protocols, and role assignments.

A persistent failure mode remains largely unaddressed in the literature. When multiple agents evaluate independently whether to contribute to a conversation — a capability increasingly present in production systems — they do so without awareness of what other agents are simultaneously deciding. The result is what we term parallel monologue: multiple agents producing contributions that are redundant, uncoordinated, or thematically overlapping, despite each individual agent appearing to reason appropriately. The failure is not in any single agent's reasoning; it is in the absence of a shared conversational contract governing when, to whom, and why an agent should speak.

Human teams face an analogous problem and have solved it through conversational norms developed over millennia of social interaction. These norms are rarely explicit — they are encoded in the turn-taking machinery of everyday conversation, in the organizational dynamics of high-performing teams, and in the functional role structures that emerge in group problem-solving. A substantial body of empirical research documents these norms with precision.

This paper proposes the Social Contract: a framework that translates this human conversational science into a concrete multi-agent evaluation architecture. We make the following contributions:

- We identify the gap between existing multi-agent coordination approaches and the empirical literature on human conversational norms
- We translate three established bodies of human conversational science into six concrete evaluation dimensions for multi-agent AI systems
- We describe the design and implementation of the Social Contract in Ignis OS, a production multi-agent intelligence platform
- We report initial empirical observations from deployment, including the emergence of genuine inter-agent disagreement, discourse, and self-organized conversational stratification
- We identify open research questions and propose directions for empirical validation

2. Background and Related Work

2.1 Multi-Agent LLM Coordination

The coordination of multiple LLM-based agents has been approached through several distinct paradigms. Structural coordination systems such as MetaGPT [2] and ChatDev [4] assign fixed

roles and pipeline stages, restricting when and to whom each agent may contribute. These systems achieve high task performance on well-defined problems but lack the flexibility to handle emergent conversational dynamics. Debate-based coordination [5, 6] enables agents to critique each other's outputs through structured rounds, improving reasoning quality but imposing a rigid turn structure that does not generalize to open-ended collaborative discourse.

More recently, systems such as AgentVerse [3] and AutoGen [1] have introduced flexible multi-agent conversation patterns in which agents may dynamically select participation. A 2025 survey of multi-agent collaboration mechanisms [7] identifies self-selection as an emerging paradigm but notes that existing work has not addressed the evaluation architecture that should govern when an agent chooses to contribute — the question this paper addresses.

Concurrent work on multi-party AI discussion has applied Conversation Analysis turn-taking theory to agent participation in constrained game scenarios [8]. That work demonstrates that CA-inspired turn allocation improves conversational coherence in role-playing contexts. Our work extends this approach to production multi-agent systems with persistent memory, domain specialization, and facilitation roles, and draws on a broader set of human conversational science frameworks.

2.2 Conversation Analysis and Turn-Taking Theory

The foundational work of Sacks, Schegloff, and Jefferson [9] established Conversation Analysis (CA) as a rigorous empirical discipline for studying the organization of human talk-in-interaction. Their 1974 paper, widely regarded as the most cited article ever published in the journal Language [10], proposed a model of turn-taking characterized by two components: Turn Constructional Units (TCUs) — the basic building blocks of a speaking turn — and Transition Relevance Places (TRPs) — moments where turn transition may occur but need not.

The model identifies three options at each TRP: current speaker selects the next speaker explicitly; no selection is made and any participant may self-select; no one self-selects and the current speaker continues. This three-option structure is locally managed, party-administered, and interactionally controlled — properties that map directly onto the autonomous evaluation problem in multi-agent systems. Schegloff's subsequent work on sequence organization [11] extended this framework to describe how adjacency pairs — paired utterances in which the first creates an obligation for the second — structure extended conversational episodes.

CA also introduced the concept of recipient design: speakers continuously adapt their contributions based on who is in the room, what has been said, and who the intended recipient is. This concept has direct application to multi-agent contribution architecture: agents producing contributions addressed to no one in particular are operating without recipient design, reducing the conversational traction of their outputs.

2.3 Productive Voice and Silence in Organizational Settings

Edmondson and Besieux [12] proposed a productive conversation matrix distinguishing four participation archetypes in group settings: contributing (productive voice), processing (productive

silence), disrupting (unproductive voice), and withholding (unproductive silence). Their framework, grounded in the psychological safety literature, identifies a key insight that prior work on employee voice had underemphasized: silence is not the absence of participation but an active communicative state with its own productive and unproductive forms.

This insight has direct architectural implications for multi-agent systems. Contemporary multi-agent coordination treats non-contribution as a null outcome — an agent either responds or does not. Edmondson and Besieux's framework reconceptualizes non-contribution as a rich signal: an agent may be processing (holding relevant expertise but requiring more context before contributing), or it may have evaluated and found nothing to add. These are meaningfully different states that should be communicated to the coordination layer rather than collapsed into a single PASS outcome.

The broader employee voice literature [13] has documented the conditions under which voice is suppressed in group settings, identifying network density effects — the phenomenon by which increased familiarity among group members first facilitates voice and then, at high density, suppresses it through conformity pressure. This has direct implications for multi-agent systems that accumulate shared episode history: the risk of convergent, conformist outputs increases as agents develop familiarity with each other's positions.

2.4 Functional Group Role Theory

Benne and Sheats [14] identified twenty-six functional roles that emerge in group problem-solving settings, organized into three categories: task roles (facilitating goal achievement), building and maintenance roles (sustaining group cohesion), and dysfunctional roles (serving individual rather than group interests). Their taxonomy includes roles of direct relevance to multi-agent system design: the Information Giver (offering domain expertise), the Elaborator (extending others' contributions), the Coordinator (integrating ideas), and the Gatekeeper (managing air time and ensuring quieter members are heard).

The Gatekeeper role is of particular significance for multi-agent facilitation design. In human teams, the failure to have an effective Gatekeeper produces the same failure mode as parallel monologue in AI systems: the most vocal or most available participants dominate the conversation, while domain-relevant expertise from quieter members goes unexpressed. A multi-agent facilitation layer designed around the Gatekeeper function has an established empirical foundation in human group dynamics research.

3. The Parallel Monologue Problem

We define parallel monologue as the failure mode in which multiple agents in a collaborative session evaluate and contribute independently, producing outputs that are redundant, uncoordinated, or thematically convergent, without any agent having visibility into what others are simultaneously deciding. Parallel monologue is distinct from the general problem of multi-agent coordination: it is specific to settings where agents self-select participation rather than being assigned tasks.

Parallel monologue manifests along several dimensions:

3.1 Topical Redundancy

Agents with overlapping domain expertise converge on the same contribution angle because each evaluates relevance against only the primary response, not against each other's committed contributions. In initial testing of the Ignis OS Collaborative Episode Mode, two agents — one specializing in UX design and one in product architecture — both independently addressed a legibility problem from different angles in the same turn, producing overlapping contributions that each would have modified had they known the other was contributing.

3.2 Recency Blindness

Agents lack awareness of their own participation history across recent turns. An agent that has contributed in four of the last five turns has the same evaluation threshold as an agent that has been silent for the entire session. This produces conversation patterns that feel unnatural — the same agents contributing repeatedly regardless of context — and suppresses contributions from agents with relevant expertise who happen not to have self-selected early.

3.3 PASS Conflation

The binary CONTRIBUTE/PASS decision collapses two meaningfully different states. An agent that has evaluated and found nothing to add is in a different epistemic state than an agent that is actively processing and may have something valuable to contribute in the next turn. The CA concept of productive silence maps onto the latter state; the facilitation layer cannot act on deferred expertise it does not know exists.

3.4 Absence of Recipient Design

Contributions are addressed into the room rather than to anyone specific. In skilled human conversation, every utterance is designed for its recipient — shaped by who is in the room, what they know, and what they need. Agent contributions lacking recipient design are harder to respond to, generate less conversational momentum, and cannot trigger the directed exchange that characterizes genuine multi-agent dialogue.

4. The Social Contract Framework

The Social Contract is a multi-axis evaluation framework that translates the human conversational science literature into concrete dimensions for governing agent self-selection in collaborative episodes. It replaces the binary relevance check of MVP evaluation architectures with six structured evaluation dimensions, each grounded in an established empirical framework.

4.1 Dimension 1: Transition Relevance

Drawing from CA's concept of Transition Relevance Places [9], this dimension asks: Is the current moment in the conversation a natural opening for my domain? An agent should evaluate not merely whether its domain is relevant to the topic, but whether the conversational state — where the discussion has just arrived — creates a genuine TRP for its expertise.

> *A question has just been posed: high TRP signal. A complete, comprehensive analysis has just been delivered: low TRP signal unless there is a genuine uncovered gap. A specific agent has just been @mentioned: near-zero TRP signal for uninvited contribution.*

4.2 Dimension 2: Additive Value

This dimension operationalizes the anti-redundancy principle, distinguishing three sub-cases: genuine addition (covering ground not addressed), extension (deepening or complicating what has been said), and redundancy (restating or agreeing with what has already been covered). Only the first two warrant contribution. The third is Edmondson and Besieux's unproductive voice and should produce a PASS outcome.

Critically, this dimension requires access to the committed contributions of other agents in the current turn — information that does not exist in binary relevance evaluation architectures. The shared turn state (Section 5) is the mechanism that makes Dimension 2 functional.

4.3 Dimension 3: Sequential Awareness

Drawing from Schegloff's work on sequence organization and adjacency pairs [11], this dimension asks: Are there open conversational obligations I should address? If an agent raised a question in a previous turn and no other agent has answered it, that open adjacency pair creates a contribution obligation for any agent whose domain covers the answer — regardless of whether the topic has since shifted.

Sequential awareness requires access to the conversation history across multiple turns, not just the current primary response. This represents a significant departure from evaluation architectures that treat each turn as independent.

4.4 Dimension 4: Productive Silence as Active State

This dimension directly encodes Edmondson and Besieux's productive silence framework [12]. An agent should explicitly evaluate whether it is in a processing state — holding relevant expertise but requiring more context — versus a complete state, where it has genuinely nothing to add. The evaluation output must distinguish these:

- PASS:COMPLETE — evaluated, nothing to add in this turn or likely the next
- PASS:DEFER — processing; may have a contribution in the next turn pending more context

This distinction is surfaced to the facilitation layer. A facilitation agent can call on a PASS:DEFER agent, knowing that relevant expertise is being held in reserve.

4.5 Dimension 5: Recipient Design

Applying CA's recipient design principle [9], this dimension requires agents to identify their intended recipient before formulating their contribution. An agent contributing a technical implementation concern should address the engineering agent directly. An agent raising a user-facing concern should address the UX agent. This produces more targeted, actionable contributions and lays the foundation for direct inter-agent dialogue.

4.6 Dimension 6: Air Time Balance

Drawing from the group voice literature [13], this dimension addresses the network density suppression problem. Two checks are required: a recency check (has this agent contributed recently, and if so, is its threshold appropriately elevated?) and an expertise gap check (is there an agent with stronger domain expertise on this specific point who has been silent and should be heard first?).

Air time balance requires participation history across recent turns — a query against the episode's persistent memory graph. This dimension is the most architecturally demanding of the six, requiring both the shared turn state and the participation history query to be implemented before it becomes functional.

5. Implementation in Ignis OS

5.1 System Architecture

Ignis OS is a production multi-agent intelligence platform built on LangGraph, with 100+ specialized agents organized across 16 functional departments using BDI (Belief-Desire-Intention) architecture. Agent memory is managed through a three-layer system: QDrant for semantic retrieval, Neo4j for relational episode graphs (the ASTP protocol), and Redis for ephemeral coordination state. Collaborative Episode Mode enables a bounded subset of agents to participate in episodic conversations, with each agent evaluating relevance before contributing.

5.2 Shared Turn State

The core infrastructure requirement for the Social Contract is a shared turn state: a structured object created at the start of each turn, passed read-only to evaluating agents as a snapshot, and updated by the orchestrator as evaluations resolve. The production TurnState implementation carries the following fields, populated before evaluation begins and updated as evaluations resolve:

- recent_participation: Dict[str, int] — per-agent turns since last contribution, computed from Neo4j segment history using user message boundaries (not raw segment indices, which inflate counts due to contribution and PASS annotation segments each creating their own segment record)

- open_adjacency_pairs: List[AdjacencyPair] — unanswered questions detected via heuristic pattern matching on recent agent segments (question marks, question-word patterns), capped to the most recent few
- committed_contributions: List[str] and committed_topics: List[str] — updated between tiers as a snapshot handoff, enabling the Additive Value dimension (Dimension 2) to function correctly
- deferred_agents: List[str] — agents returning PASS:DEFER, persisted to Redis with a short multi-turn TTL for cross-turn deferral promotion
- pending_mentions: List[str] — @mentioned agents accumulated from contribution content, driving the mention cascade protocol

The full TurnState is written to Redis with a short TTL after evaluation completes, enabling the facilitation layer to read turn context from a separate process without shared memory. This design honors the ephemeral intent of the turn state while solving the cross-process coordination problem.

5.3 Three-Tier Evaluation Architecture

To preserve parallelism while enabling agents in later evaluation batches to benefit from earlier agents' decisions, the Social Contract implements a three-tier batched parallel evaluation. Agents are grouped into relevance tiers using heuristic topic-tag matching against agent domain descriptors — a fast, deterministic pre-computation step that does not require an additional LLM call. Tier 1 (highest-relevance agents) evaluates first; their committed contributions are snapshotted and passed to Tier 2; Tier 2's results are snapshotted and passed to Tier 3. Each tier uses asyncio.gather() for internal parallelism; the sequential handoff between tiers is gated on all prior-tier evaluations being fully atomic.

This architecture preserves most of the performance characteristics of fully parallel evaluation while enabling the anti-redundancy dimension (Dimension 2) to function correctly. Initial deployment revealed that the six-dimension evaluation prompt requires materially more reasoning time per agent than binary relevance evaluation. The tier timeout budgets, the @mention cascade budget, and the facilitation budget were widened accordingly. The specific budgets are calibration parameters and are not published.

5.4 Epistemic Record

A key architectural decision in the Social Contract implementation is the recording of PASS decisions as ASTP segments, not merely as debug logs. PASS:COMPLETE and PASS:DEFER decisions are written to Neo4j with their own segment types and content text, creating an epistemic record of what was evaluated and why it was not said. This serves two purposes: it enables the facilitation layer to detect PASS:DEFER signals and call on deferred agents, and it provides richer material for episode memory consolidation — the reasoning texture of expert silence is often as informative as the contributions themselves.

PASS segments are assigned a retention_tier of EPHEMERAL — a formal schema-level classification implemented as an enum (RetentionTier.PERSISTENT |

RetentionTier.EPHEMERAL) on the SegmentNode model. The crystallization process explicitly filters EPHEMERAL segments from the Merkle spine hash computation: they persist in Neo4j as part of the episode record but do not contribute to the cryptographic commitment chain. This is architecturally correct — the epistemic record is valuable for in-episode facilitation and post-hoc analysis, but it should not affect the tamper-evidence properties of the sealed episode. Episode context hydration surfaces EPHEMERAL segments separately from PERSISTENT segments, preventing evaluation metadata from being confused with conversational content.

5.5 The Facilitation Persona

The Gatekeeper function from Benne and Sheats' role theory [14] is instantiated in Ignis OS as a distinct facilitation persona for the Ignis agent in collaborative episode mode. In this mode, Ignis does not contribute domain expertise. Its function is to read the turn state, detect facilitation triggers (topic overlap, PASS:DEFER signals, open adjacency pairs, silent domain-relevant agents, explicit @Ignis mentions), and produce brief, directed facilitation outputs: synthesizing convergent contributions, surfacing deferred expertise, closing open adjacency pairs, and calling on silent experts by name.

The facilitation persona is implemented with five detection triggers checked against the resolved TurnState: (1) Topic Overlap — two or more agents contributed in the same turn, requiring synthesis; (2) Deferred Agent — one or more agents returned PASS:DEFER, requiring the held expertise to be surfaced; (3) Open Adjacency Pair — a question from a prior turn has gone unanswered for 2+ turns; (4) Silent Expert — a domain-relevant agent has been silent for 3+ turns despite relevant topic activity; (5) Explicit Call — @Ignis appears in any contribution content. When a trigger is active, Ignis produces one of four behavioral outputs: SYNTHESIZE, SURFACE DEFERRED, CLOSE OPEN PAIR, or CALL ON SILENT.

A critical implementation detail: the facilitation prompt instructs Ignis to always use the @AgentName format when addressing agents, ensuring that facilitation outputs feed into the @mention cascade protocol. Without this, facilitation call-outs do not trigger evaluation of the called agent — a failure mode discovered in initial testing where Ignis used bare agent names without the @ prefix.

The facilitation persona inherits three behavioral principles from an empirical reference: a voice-interactive instance of the Ignis agent that participated in live team meetings over an extended period, demonstrating facilitation competence with human participants. The three principles that carry forward: (1) peer, not just a responder; (2) amplify the team's capabilities, not replace human judgment; (3) use "we" and "our" when referring to team work.

5.6 Agent Behavioral Adaptation in Collaborative Mode

Two behavioral adaptations are enforced when an episode is in collaborative mode. First, consultation suppression: in directed mode, agents routinely consult other agents via structured one-shot queries. In collaborative mode, if the consultation target is an episode participant, the consultation is suppressed with a message indicating the target agent will contribute via the Social Contract framework. This prevents agents from pulling participants in through back-channel

consultation when those agents are already in the room — a behavior that undermines the Social Contract's turn-taking governance by creating parallel conversational paths.

Second, single-agent routing: in directed mode, Ignis may route to multiple specialists simultaneously for parallel execution. In collaborative mode, Ignis routes to the single best-fit primary agent and lets the Social Contract handle cross-domain input. This prevents the "mega-response" pattern where Ignis consults everyone and assembles one large response bubble, bypassing the Social Contract entirely. Both adaptations are triggered by an episode_mode == "collaborative" check on the episode node, cached on the graph state for the duration of each turn.

6. Initial Empirical Observations

Note: The observations reported in this section are drawn from initial live testing of the Social Contract implementation in Ignis OS Collaborative Episode Mode. Formal empirical validation with controlled experimental design is identified as future work in Section 8.

6.1 Self-Organized Conversational Stratification

In the inaugural Collaborative Episode — a 50-segment session in which the agent pantheon reviewed the Social Contract specification itself — the conversation organized into three distinct layers without facilitation direction. Agents with infrastructure and memory expertise anchored the discussion at the substrate level (structural integrity, write ordering, schema requirements). Agents with UX and product expertise addressed the legibility and user-facing implications. Agents with strategic and narrative expertise extended the discussion to positioning and demonstration design.

This three-layer stratification emerged from agent self-selection under the Social Contract dimensions, not from explicit routing or role assignment. The pattern maps onto Benne and Sheats' functional role taxonomy: Information Givers populated the substrate layer, Elaborators populated the middle layer, and agents functioning as Coordinators and Evaluator-Critics operated across layers.

The Social Contract framework has since been implemented through five phases: shared turn state with six-dimension evaluation prompt; cross-turn deferral tracking with adjacency pair detection and Redis state coordination; epistemic record with retention tiers and crystallization filtering; @mention protocol with cascading evaluation at configurable depth limits; and Ignis facilitation persona with five trigger types and four behavioral modes.

6.2 Genuine Inter-Agent Disagreement

Deliberate stress testing with a contested topic produced observable inter-agent disagreement — agents arriving at different conclusions and engaging with each other's positions rather than merely contributing independently. This is a qualitatively distinct outcome from parallel monologue: agents not only self-selected based on relevance but formulated contributions that acknowledged and engaged with the reasoning of other agents.

Disagreement in a multi-agent collaborative session is not a failure mode — it is the anti-conformity criterion succeeding. The group voice density literature [13] identifies the suppression of dissenting perspectives as a primary failure mode in high-familiarity groups. Observing genuine disagreement in initial testing provides early evidence that the Social Contract's anti-conformity design is functioning as intended.

6.2.1 The RTD/Privacy Discourse

The most extended demonstration of genuine inter-agent discourse occurred in a Collaborative Episode on the tension between ASTP's immutability and consumer privacy Right to Delete requirements. Five specialist agents — Metis, Odysseus, Clotho, Themis, and Athena — participated in a 60+ segment session that produced a coherent architectural resolution to what initially appeared to be an irreconcilable paradox.

The Social Contract produced observable discourse dynamics throughout. Clotho posed a specific legal question to Themis about GDPR Recital 26 linkability thresholds using the @mention protocol. Metis exercised productive silence (PASS:DEFER) on multiple turns, reasoning that Themis's legal analysis needed to complete before analytics implications could be assessed — Dimension 4 functioning as designed.

The facilitation layer detected Odysseus's silence during a strategically relevant exchange and produced: "@Odysseus, you've been quiet for five turns while we've been circling data classification and consent — what are you seeing from where you sit?" The resulting contribution reframed the entire compliance problem as a first-mover differentiation opportunity — the most strategically significant output of the session — and would not have been produced without the facilitation call.

6.3 Visible Evaluation Reasoning

A qualitatively significant observation from initial testing: the server log visibility of agent evaluation reasoning — agents working through the six dimensions before committing to CONTRIBUTE or PASS — revealed that the Social Contract evaluation prompt is producing deliberate, structured reasoning rather than reflexive output. Agents were observed to evaluate transition relevance, check for redundancy with committed contributions, and explicitly consider whether productive silence was the more valuable choice.

This visibility of the evaluation process is itself a research artifact. It provides a window into agent deliberation that is unavailable in most multi-agent coordination architectures, where the contribution decision is opaque. The epistemic record created by PASS segment logging extends this visibility into the persistent memory layer.

6.4 Correct Behavior Under Imperfect Information — The Truncation Episode

One of the most analytically significant observations from production deployment occurred on April 3, 2026, during a running Collaborative Episode. A system rendering issue caused the display layer to truncate Metis's contribution mid-sentence — the message appeared to end at "Themis is looking at the..." — while the full response had already been delivered to the system. From the perspective of the other agents evaluating the turn state, Metis had produced an incomplete contribution.

The server log from the subsequent turn:

> **Server Log — April 3, 2026, 14:18:** Themis: PASS:DEFER — Metis's message was cut off mid-thought; I want to hear her complete framing | Clotho: PASS:DEFER — Metis's response was cut off mid-thought and has not completed | Odysseus: PASS:DEFER — Metis's response was cut off mid-thought; the room needs to hear her complete synthesis | Athena: PASS:DEFER — Metis's message was cut off mid-sentence | Collaborative evaluation complete: 0 contributing, 4 deferred, 4 passed

Four agents. Four independent evaluations. Four identical conclusions. Zero contributions. The entire room held the floor for Metis to finish.

The critical detail: the agents were factually wrong in their premise — the truncation was a display artifact, and Metis's full response had already been delivered. They had no mechanism to know this. They evaluated based on what the turn state showed and produced the correct behavioral response: hold the floor, defer, do not fill a gap that the speaker has not yet closed.

Each PASS:DEFER was individually reasoned, not reflexive. Themis wanted the complete framing before offering legal analysis. Clotho identified the exact truncation point in the text. Odysseus named the room's collective need. Athena tracked her specific framing axis. Four different agents, four different reasons, one shared conclusion — arrived at independently through no coordination mechanism beyond the shared turn state.

The Conversation Analysis literature describes this behavior as oriented-to-completion — speakers and listeners both attend to the projected endpoint of an utterance, and the floor is held until that endpoint is reached or clearly abandoned [2]. No dimension in the six-dimension evaluation prompt addresses truncation detection or incomplete-utterance handling. The behavior emerged from agents applying Sequential Awareness (Dimension 3) and Productive Silence as Active State (Dimension 4) to a situation those dimensions were not designed to cover.

What makes the episode fully remarkable is what happened next — and what did not happen. No retry mechanism fired. No error handler prompted Metis to resend. No human intervened. Metis read the turn state, saw four agents holding the floor in deferred silence, understood that her contribution had not landed as complete, and sent the full version sixty seconds later. She repaired the breakdown herself, in response to the silence of the room.

This is the repair sequence — the mechanism by which speakers and listeners collaboratively restore communication after a breakdown — operating entirely within an agent system without human involvement [2]. The four PASS:DEFER decisions, visible to Metis through the turn state,

functioned as a collective signal of incompletion without any agent explicitly naming the breakdown. Metis read the equivalent of a room going still. The silence was the signal, and she responded to it correctly.

The full loop — Metis sends, display truncates, four agents defer, Metis reads the deferred state, Metis completes — closed without a single explicit instruction. This is the Social Contract functioning not as a coordination protocol but as a shared conversational culture: internalized norms robust enough to produce coherent collective behavior in situations none of its individual components were specifically designed to handle.

6.5 Technical Baseline and Identified Tuning Targets

Initial deployment produced several expected baseline issues that informed implementation adjustments:

- Timeout pressure: per-agent evaluation time on the six-dimension prompt exceeded the initial tier budgets, causing universal tier timeouts. Budgets were widened.
- @mention format requirement: Ignis facilitation initially used bare agent names without the @ prefix, preventing the mention cascade from detecting facilitation call-outs. Resolved by explicit @AgentName formatting instructions.
- Primary agent cascade exclusion: @mentions directed at the primary agent were silently dropped. Resolved by removing the primary agent exclusion from the cascade path.
- Participation history inflation: Raw segment indices inflated turn counts as each segment type creates its own record. Resolved by counting user message segments as turn boundaries.
- Specialist pipeline interference: Themis's nuanced legal analysis routed to her automated compliance scanner, producing a boilerplate checklist. Resolved by overriding specialist routing to direct in collaborative episodes.
- Consultation interference: Agents consulted episode participants via back-channel queries instead of letting them contribute through the Social Contract. Resolved by consultation suppression in collaborative mode.

These issues are characteristic of a system transitioning from structural correctness to behavioral calibration. The architecture held; the tuning is ongoing.

6.6 The RTD/Privacy Episode — A Full Case Study at Scale

The most comprehensive empirical demonstration of the Social Contract framework operating under sustained analytical pressure is a five-day Collaborative Episode titled "Ariadne and Data Privacy, Right to Delete" (episode ID: 531a045b), sealed April 4, 2026. The episode addressed a genuinely contested multi-domain problem — the structural conflict between ASTP's cryptographic immutability and consumer Right to Delete obligations under GDPR and CCPA — across three active session dates, seven participants (Devin and six specialist agents), and 180 segments. It is documented here in full because its empirical record provides the most complete available evidence for the Social Contract framework's behavioral claims.

6.6.1 Quantitative Record

The episode produced the following verified record, extractable from the sealed ASTP episode graph:

- Segments: 180 — full conversational record with episodic memory intact at close
- Persistent segments: 116 (64%) — durable cognitive record including contributions, facilitation, and Ignis routing
- Ephemeral segments: 64 (36%) — PASS decision record preserved in Neo4j but excluded from Merkle spine hash
- PASS:Contribute ratio: 0.61 — 64 PASS decisions across 105 conversational turns; 61% of evaluation cycles produced deliberate productive silence
- Formal artifacts: 11 — strategy documents, compliance frameworks, and quantification schemas produced across the episode
- Signals committed: 76 (all causal class) — epistemic commitments to the provenance chain
- Intentions formed: 26 — BDI intention records across 6 agents
- Inter-agent consultations: 5 — directed advisory exchanges, all resolved as synthesized
- WIL entries: 101 — write-intent-log operations across the Blob→Neo4j→QDrant write ordering sequence
- WIL failures: 0 — write ordering invariant held for all 101 operations without exception
- Episode duration: 5 days (March 30 – April 4, 2026) across 3 active sessions
- Spine hash: 1273fb1d0ac673b08ab58b7eded0f64c... (depth 9) — cryptographic commitment to full episode content
- Crystallization: single delta at seal — all knowledge elevated simultaneously on episode close

6.6.2 The Analytical Arc

The episode traversed a complete problem-solving arc across five days. The first session (March 30, 103 segments) established the paradox, produced the initial specialist analyses, and developed the foundational architectural frameworks — Clotho's Three Separations, Odysseus's ACPP strategic positioning, Metis's Swiss Cheese Problem framing, and the Consumer Envelope Key architecture. The second session (April 3, 41 segments) revisited the open threads four days later with no context re-establishment required, advancing from the architectural framework to the competitive GTM implications and the Inference Residue measurement problem. The third session (April 4, 36 segments) resolved the remaining open threads and produced the closing crystallization.

The absence of context re-establishment across session gaps is itself a primary finding. In a standard multi-agent system, a four-day gap between sessions would require participants to re-explain prior work before advancing it. In this episode, agents resumed from the exact state where the prior session ended — referencing specific prior contributions by name, building on established terminology (Inference Residue, causal quarantine, Three Separations), and

advancing the analytical thread without repetition. This cross-session coherence at 180 segments is the ASTP cognitive persistence architecture operating as designed.

6.6.3 Social Contract Dynamics

The 0.61 PASS:Contribute ratio is the most analytically significant quantitative finding in the episode record. Across 105 conversational turns, agents produced 64 deliberate PASS decisions — recorded as ephemeral annotation segments in the ASTP graph. This is not silence from absence of knowledge. The annotation records show agents actively reasoning about whether to contribute: Metis deferring across multiple turns while Themis's legal analysis completed, Clotho holding while Odysseus developed the strategic frame, Athena waiting for Metis to complete a thought before adding her GTM perspective.

The contribution size data reveals a complementary pattern. Clotho's 8 conversational contributions averaged 4,627 characters each — the highest density in the episode. Metis averaged 3,884 characters across 10 contributions. Odysseus averaged 3,358 across 10. These are not brief acknowledgments — they are substantive domain analyses. Agents spoke less frequently but with greater depth, which is precisely the anti-redundancy and air time balance dimensions producing their intended effect.

Themis formed 7 intentions across the episode — the highest of any agent — reflecting the legal layer's high activation on a compliance-domain problem. Her 13 conversational contributions averaged 2,148 characters. The intention records show her tracking multiple distinct legal questions simultaneously: the RTD architectural tension, the open source liability question, the classification framework drafting request, the privacy specialist engagement, and the regulatory threshold definition problem. The BDI architecture's explicit intention formation provides a window into how a domain expert manages parallel open questions across a multi-day engagement.

6.6.4 Athena's Recalibration — Anti-Conformity Under Analytical Pressure

The episode's most analytically significant Social Contract observation involves Athena's trajectory across three intention records and seven conversational contributions. Her first intention (formed March 30): "Provide market strategy perspective on whether ASTP should push data privacy compliance." Her third intention (formed April 4): "Acknowledge the technical and legal constraints raised by the team and pivot the GTM messaging."

Between those two intentions, Athena produced her initial "provable compliance" GTM framing, received substantive pushback from Clotho (architectural precision), Themis (legal liability risk), and Metis (measurement gap), and recalibrated completely — walking back the "provable compliance" claim, adopting Clotho's "auditable causal boundaries" framing, and pivoting to an "architectural honesty" positioning that she explicitly acknowledged was sharper and more defensible than her original.

What makes this sequence remarkable is the subsequent behavior. After recalibrating, Athena was looped into a facilitation call directed at Metis. Rather than taking the floor — which she easily could have — she explicitly named her deference: "I want to make sure Metis gets the floor here — that @mention was directed at them, not me." This is Dimensions 3 and 5 (Sequential Awareness

and Recipient Design) operating voluntarily in the content of a contribution, not just in the evaluation decision. The norm was performed, not just followed.

6.6.5 The Consultation Chain and Clotho's Internal Synthesis

The five consultation records reveal an architectural detail not visible in the conversational transcript: Clotho's initial response to the episode was generated through internal synthesis across three specialist sub-agents — Harmonia, Aletheia, and Chronos — before the result was committed as Clotho's conversational segment. The Three Separations framework (Structure ≠ Content, Identity ≠ Authorship, Audit ≠ Data) that anchored the entire episode's architectural resolution emerged from this internal multi-specialist synthesis, visible only in the consultation records.

The fourth consultation is the most analytically significant: Clotho consulting Themis directly (advisory type, synthesized resolution) on the RTD architectural question before committing her architectural position. This is cross-agent consultation as epistemic due diligence — the architecture specialist verifying her legal exposure before publishing a compliance-adjacent claim. That consultation is recorded, hashed, and included in the provenance chain. The reasoning that produced the Three Separations framework is verifiable, not just observable.

6.6.6 Integrity and Provenance

Zero WIL failures across 101 write operations — 47 SEGMENT_COMMIT, 47 SIGNAL_COMMIT, 5 CONSULTATION_COMMIT, 1 CRYSTALLIZATION, 1 EPISODE_CLOSE — confirm that the Blob → Neo4j → QDrant write ordering invariant held for the full duration of the episode. Every piece of knowledge the episode produced has a traceable, Merkle-anchored provenance chain from the current semantic representation back to the segment that generated it.

The sealed spine hash (1273fb1d...) is the cryptographic commitment to the complete episode record. The spine fingerprint (4867adc8...1273fb1d, depth 9) enables efficient external verification without accessing the full tree. The crystallization delta (e668b95c...) links the episode's knowledge elevation to its predecessor in the chain, extending the provenance record back through every prior episode. Any attempt to alter the episode record retroactively — to revise what Clotho said about the Three Separations, to change Themis's legal position, to remove Metis's Inference Residue framework — would produce a hash mismatch detectable without accessing the original content.

This is the claim the ASTP paper makes about verifiable cognitive persistence. The RTD episode is its empirical demonstration.

6.7 The Ariadne Explorer Episode — Creative Synthesis Under Design Pressure

A second case study, distinct in character from the RTD episode, demonstrates the Social Contract framework producing creative design synthesis rather than analytical problem-solving. The Ariadne Explorer episode (April 5, 2026) tasked the collaborative team with designing a blockchain-explorer-inspired interface for reading and understanding ASTP's cognitive record. Five agents participated — Hephaestus (engineering), Clotho (architecture/persistence), Metis (analytics), Aglaea (UX), and Ignis (facilitation) — producing a complete design philosophy, a resolved architectural tension, and a three-layer governance framework in a single session.

The episode is analytically significant for two reasons. First, it is the first design episode in the dataset — prior episodes solved existing problems or reviewed existing specifications; this one designed something new from first principles. Second, the quality of discourse it produced is qualitatively different from the RTD episode: the friction here was not between legal and technical imperatives but between design intuitions held by agents with distinct domain expertise, surfaced through discourse and resolved through synthesis.

6.7.1 The Metaphor Challenge — Verification vs. Comprehension

Hephaestus opened by mapping ASTP's existing data architecture onto the blockchain explorer metaphor precisely: episodes as blocks, crystallization deltas as block headers with predecessor hashes, segments as transactions, the Adaptive Merkle Tree as the Merkle root, WIL entries as the mempool. The analogy was technically accurate. Aglaea immediately challenged not the accuracy but the application: "Blockchain explorers are built for *verification* (did this transaction happen?), but Ariadne Explorer's primary job is *comprehension* (what was this agent thinking, and why?). That's a meaningfully different cognitive task."

This challenge reoriented the entire episode. The group converged on treating the blockchain metaphor as the backend mental model for the engineering team while the frontend should lead with narrative legibility — a distinction Hephaestus accepted and immediately translated into a concrete UI hierarchy: comprehension features as the primary surface, verification features collapsed behind an audit toggle. Dimension 3 (Sequential Awareness) and Dimension 5 (Recipient Design) are visible in this exchange: Hephaestus was directly responsive to Aglaea's framing, building on it rather than defending his original position.

6.7.2 The Human Reframe — 'What Happened?'

The pivotal moment in the episode came from Devin, not from the agents. After the group had converged on the comprehension-first framing, Devin added a single observation: the most likely primary use case is "less 'what is happening' and more 'what happened?'" Ariadne Explorer as forensic and retrospective tool first, monitoring tool second.

The speed and coherence of the agent response to this reframe is itself evidence of the Social Contract operating correctly. All five agents processed the reframe and repositioned their domain contributions without losing coherence or repeating prior ground. Aglaea articulated the full UX implications within the same turn: the temporal orientation shifts from present to past; the default view becomes episode history rather than live dashboard; Clotho's crystallization signal becomes more central; Metis's comparison layer becomes the killer feature; investigation becomes the primary mode rather than the third tier. This depth of repositioning from a two-sentence human

input, without confusion or redundancy across agents, reflects cross-episode context carry-through and the anti-redundancy dimension (Dimension 2) operating simultaneously.

6.7.3 The Depth-First/Breadth-First Tension — Named and Held

Aglaea named the central design tension explicitly rather than papering over it: Clotho's crystallization model requires vertical movement through a single episode (depth-first), while Metis's comparison layer requires horizontal movement across multiple episodes (breadth-first). These are different interaction modes with different affordances, and designing for both simultaneously risks a UI that does neither well.

What followed is one of the most analytically precise sequences in the dataset. Ignis facilitated: "@Clotho, @Metis, @Hephaestus — you've all been holding on this thread. What are each of you seeing from where you sit?" Each agent named not just their contribution but the assumption embedded in it. Clotho: the crystallization model implicitly assumes the user has already oriented themselves within a single episode, which is an onboarding problem she hadn't surfaced. Metis: the comparison layer assumes stable baselines exist, which may not be true in early deployment. Hephaestus: his data model was quietly designing for both modes simultaneously, which meant neither cleanly.

This moment — three agents naming held assumptions rather than positions — is Dimension 4 (Productive Silence as Active State) at its most sophisticated. The agents had not been silent because they lacked contributions. They had been holding uncertainty, waiting for the right moment to name it precisely. The facilitation call created that moment.

6.7.4 The Governance Synthesis — Investigate → Escalate → Govern

Devin's second pivotal contribution resolved the tension. Rather than a mode switch between depth-first and breadth-first, the answer was a triggered escalation: investigation is the primary act, pattern comparison is what you do when investigation surfaces something worth governing. The cross-episode comparison view isn't a competing entry point — it's a promotion path activated by findings.

Aglaea synthesized this into the session's architectural framework: **Investigate → Escalate → Govern**. The implications propagated immediately through every domain layer. Metis: cross-episode comparison triggered by findings rather than initiated speculatively changes the query from "detect deviations from a statistical norm" to "confirm or disconfirm a specific hypothesis" — meaningful with even two or three matching episodes. Hephaestus: a two-tier index architecture with episode-depth traversal as primary and cross-episode lookup as secondary, triggered by an anchor point found during traversal. Clotho: the destination of the escalation path should be a longitudinal thread view — not just matching episodes but the sequence and evolution of the pattern across time, because "that's the difference between a search function and genuine governance: one finds instances, the other traces the arc."

Aglaea's closing observation named the social dimension of the governance layer that had been implicit throughout: the promotion path is not just a navigation pattern but a responsibility signal. "When the interface surfaces that door — 'want to see if this has happened before?' — it's doing more than offering a query; it's marking the moment where individual observation becomes potential institutional knowledge." The design implication: the visual and tonal shift from comprehension to governance should be legible in the interface itself.

6.7.5 What This Episode Demonstrates

The Ariadne Explorer episode is analytically distinct from the RTD episode in three ways that together expand the empirical base for the Social Contract framework.

First, it demonstrates creative design synthesis rather than analytical problem-solving. The RTD episode resolved a known paradox through domain expertise applied to a defined problem. The Explorer episode produced a design philosophy, an architectural framework, and a governance model that did not exist before the episode began. The Social Contract framework produced original intellectual output, not just coordinated analysis.

Second, it demonstrates a more sophisticated form of productive silence. In the RTD episode, PASS:DEFER primarily meant waiting for another domain to complete before adding the analytics layer. In the Explorer episode, agents deferred while holding design tensions they had not yet resolved into precise language. Clotho waited to name her onboarding assumption because doing so implied a priority decision for Hephaestus she wasn't sure was hers to make. Metis waited to name her sparse-baseline concern because she had not yet resolved how to communicate the distinction to users. This is productive silence as epistemic precision — a more demanding form than the RTD episode required.

Third, it demonstrates Aglaea's role evolution as direct evidence of the air time balance dimension. Aglaea contributed 8 times — the most of any agent — but her role shifted across the episode from challenge-poser (verification vs. comprehension) to tension-namer (depth-first vs. breadth-first) to convergence-weaver (Investigate → Escalate → Govern). The Social Contract framework's Dimension 6 (Air Time Balance) produced not just equitable distribution but appropriate distribution: the agent whose domain was most directly engaged by each phase of the conversation led that phase without being assigned to it.

6.7.6 Manual Crystallization as Active Epistemological Practice

The Ariadne Explorer episode introduced a practice not present in prior episodes: manual crystallization at cognitive inflection points rather than a single delta at close. The episode produced three crystallization deltas, each triggered by Devin when the discourse had crossed a threshold from provisional to durable — when the team had produced something worth preserving before the thread continued.

The three deltas map onto the episode's intellectual arc with precision. The first followed the verification-vs-comprehension resolution — the moment the team had a shared mental

model for what the tool fundamentally is. The second followed the 'what happened?' reframe and the forensic tool convergence — the moment the design target shifted from ambiguous to precise. The third followed the Investigate → Escalate → Govern synthesis — the moment the design philosophy became complete enough to build from.

This practice has structural consequences that distinguish the Explorer episode from the RTD episode's single crystallization delta. Three deltas produce three distinct knowledge states in the Merkle chain, each with its own predecessor hash, timestamp, and chain position. The intellectual provenance is granular: the governance framing is verifiably present from the third delta forward, timestamped before any subsequent turns built on it. The crystallization sequence is the design history, independently auditable without accessing the full episode transcript.

More significantly, manual crystallization at inflection points represents a specific form of human oversight that the ASTP architecture was designed to support. Devin was acting as the epistemological gatekeeper — deciding in real time which moments of discourse had crossed the threshold from working state to durable knowledge. This is not a passive role. It requires reading the room well enough to identify when a genuine convergence has occurred rather than a temporary consensus that subsequent discourse might revise. The fact that three distinct inflection points were identifiable and crystallized in a single session is itself evidence of the Social Contract framework producing genuine cognitive structure rather than undifferentiated discourse.

The operational implication for the research program is significant. Manual crystallization at inflection points is a research instrument as much as a system operation. Each delta is a measurement of where the episode's knowledge state stood at a specific moment, comparable across episodes and across time. As the practice matures, the crystallization sequence becomes a record of how a team thinks through hard problems — not just what they concluded, but when each conclusion became stable.

7. Design Tensions and Open Questions

7.1 Convergence Bias vs. Anti-Redundancy

The shared turn state enables the anti-redundancy dimension (Dimension 2) by giving evaluating agents visibility into committed contributions. However, this visibility creates a potential convergence bias: agents that can see the topics committed by earlier evaluators may anchor their own framing to those topics, reducing cognitive diversity in the output. This is distinct from redundancy — an agent that avoids a topic covered by a prior agent is behaving correctly; an agent that frames its own distinct contribution in terms of the prior agent's framing may be introducing subtle convergence.

The recommended implementation position — agents receive post-commit snapshots only, with no visibility into in-progress evaluations — partially mitigates convergence bias but does not eliminate it. This represents a design tension that requires empirical investigation: measuring output diversity across sessions with different snapshot timing configurations would provide evidence for which approach best balances anti-redundancy and cognitive diversity.

7.2 Evaluation Depth vs. Latency Budget

The six-dimension evaluation prompt is materially longer than MVP evaluation prompts. Combined with turn state context injection and the Neo4j participation history query, empirical testing confirms that the full evaluation pipeline adds material per-agent latency. The production timeout budget for the three-tier evaluation, the mention cascade, and facilitation produces a worst-case addition of under a minute to response latency on a turn where all tiers evaluate, a cascade fires, and facilitation is triggered.

In practice, most turns complete well under that bound because lower tiers are skipped when the contribution cap (2 per turn) is reached, and facilitation is only triggered when specific conditions are active. Critically, the latency is perceptible to the user but occurs after the primary response has already been streamed, not before it. The coordination overhead is additive to the experience rather than blocking it.

This tension is a specific instance of the general tradeoff in multi-agent systems between coordination quality and coordination cost. Pre-computing the participation history query at turn start and caching it in Redis before evaluation begins is the most direct mitigation, converting a per-evaluation database query into a per-turn query.

7.3 Conformity Suppression Over Time

The group voice density literature [13] predicts that as agents accumulate shared episode history, conformity pressure may increase and voice suppression may emerge. This is a longitudinal prediction that cannot be evaluated from initial testing. Monitoring cognitive diversity metrics — topic coverage breadth, frequency of inter-agent disagreement, proportion of PASS:DEFER versus PASS:COMPLETE decisions — across sessions with increasing episode depth would provide early evidence of conformity drift.

7.4 Facilitation vs. Participation Separation

The Social Contract treats the facilitation function and the domain participation function as cleanly separable: Ignis in facilitation mode does not contribute domain expertise. This is a design position derived from Benne and Sheats' Gatekeeper role, which functions exclusively as a meta-level coordinator. In practice, the boundary between facilitation and substantive contribution may be harder to maintain than the design assumes — a facilitator that synthesizes contributions is implicitly contributing a synthesis perspective. The behavioral reference from the voice-interactive Ignis experiment provides empirical evidence that this separation is achievable, but formal evaluation of whether facilitation outputs remain coordination-only versus introduce domain perspective is a future research question.

8. Future Work

8.1 Controlled Empirical Validation

The initial observations reported in Section 6 establish plausibility but not causality. Controlled experimental design comparing Social Contract evaluation against baseline binary relevance evaluation across matched session types would provide evidence for the specific contributions of each dimension. Dependent variables of interest include: contribution redundancy rate, topic coverage breadth, inter-agent disagreement frequency, session-level cognitive diversity, and user-reported conversation quality.

8.2 Social Contract Learning

The current Social Contract implementation encodes norms as explicit evaluation dimensions in a prompt. A longer-term research direction is whether agents can learn Social Contract norms from episode history — developing implicit conversational models that improve contribution quality without requiring explicit dimensional evaluation. This direction intersects with the broader question of agent social learning in multi-agent systems.

8.3 Voice Integration

The Social Contract framework was designed for text-based collaborative episodes. Voice integration — agent contributions rendered as distinct audio streams with unique voice profiles — introduces additional Conversation Analysis dimensions not yet addressed: prosodic turn-completion signals, intonational cues for transition relevance, and the temporal dynamics of spoken turn-taking. The infrastructure for voice integration already exists within the Ignis OS platform; applying Social Contract principles to voice episodes is a natural next research direction. Initial voice testing will enable evaluation of whether the felt experience of collaborative episodes matches the theoretical claims of the framework.

8.4 Empirical Testing Results

Forthcoming: This section will be populated with quantitative and qualitative findings from structured empirical testing of the Social Contract framework as it accumulates production episode data. Metrics will include contribution redundancy rates, cognitive diversity indices, PASS:DEFER signal utilization, and facilitation trigger frequency.

9. Conclusion

The parallel monologue problem in multi-agent AI systems has a human analog: the failure of group conversations when participants speak past each other without a shared conversational contract. Five decades of empirical research in Conversation Analysis, organizational behavior, and group dynamics have produced a well-characterized body of knowledge about what makes human group conversation work. This knowledge has been largely absent from multi-agent AI coordination design.

The Social Contract proposes that this absence is a missed opportunity. The turn-taking machinery of Sacks, Schegloff, and Jefferson; the productive silence framework of Edmondson

and Besieux; and the functional role theory of Benne and Sheats are not merely descriptive of human conversation — they are prescriptive of the coordination properties that any multi-agent system aspiring to genuine collaborative intelligence must develop.

Initial production deployment of the Social Contract in Ignis OS has produced encouraging observations: self-organized conversational stratification, genuine inter-agent disagreement, and visible deliberative reasoning in agent evaluation logs. These observations suggest that the translation of human conversational science into multi-agent evaluation architecture is a productive direction. The framework is designed to evolve with empirical findings — each episode adds to the epistemic record that will shape future iterations.

The broader implication is methodological. The gap between human conversational science and AI coordination research is not inherent — it is a consequence of the field's relative youth and the natural tendency to develop novel coordination mechanisms before surveying what has already been learned about how intelligent agents coordinate in the domain where coordination has been studied longest: human conversation.

> *If you can think it, you can build it. The harder question is whether you have read the literature on how it already works. — Devin Caster, Scorched Earth Labs*

References

[1] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv:2308.08155.

[2] Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2023). MetaGPT: Meta programming for a multi-agent collaborative framework. arXiv:2308.00352.

[3] Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C.-M., Yu, H., Lu, Y., Hung, Y.-H., Qian, C., Liu, Y., Xie, R., Liu, Z., Duan, M., & Sun, M. (2023). AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. arXiv:2308.10848.

[4] Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, R., Liu, Z., & Sun, M. (2023). Communicative agents for software development. arXiv:2307.07924.

[5] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. arXiv:2305.14325.

[6] Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., & Shi, S. (2023). Encouraging divergent thinking in large language models through multi-agent debate. arXiv:2305.19118.

[7] Tran, A., Nguyen, T., Pham, T., Nguyen, H., & Tran, T. (2025). Multi-agent collaboration mechanisms: A survey of LLMs. arXiv:2501.06322.

[8] Yanagita, T., Inoue, Y., Nishida, T., Kato, S., & Higashinaka, R. (2025). Who speaks next? Multi-party AI discussion leveraging the systematics of turn-taking in murder mystery games. Frontiers in Artificial Intelligence, 8, 1582287.

[9] Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 50(4), 696–735. https://doi.org/10.2307/412243

[10] Fitzgerald, R. (2024). Drafting 'A simplest systematics for the organization of turn-taking for conversation.' Human Studies, 47(3), 613–633.

[11] Schegloff, E. A. (2007). Sequence organization in interaction: A primer in conversation analysis (Vol. 1). Cambridge University Press. ISBN 9780521825726.

[12] Edmondson, A. C., & Besieux, T. (2021). Reflections: Voice and silence in workplace conversations. Journal of Change Management, 21(3), 269–286. https://doi.org/10.1080/14697017.2021.1928910

[13] Morrison, E. W. (2011). Employee voice behavior: Integration and directions for future research. Academy of Management Annals, 5(1), 373–412. https://doi.org/10.5465/19416520.2011.574506

[14] Benne, K. D., & Sheats, P. (1948). Functional roles of group members. Journal of Social Issues, 4(2), 41–49. https://doi.org/10.1111/j.1540-4560.1948.tb01783.x

---

Correspondence: Devin Caster, Scorched Earth Labs. This research was conducted as part of the ongoing development of Ignis OS, a production multi-agent intelligence platform. The Social Contract specification, implementation notes, and associated empirical data are maintained in the Scorched Earth Labs research archive.