Constitutional Orthogonality

Anonymous

February 6, 2026

Abstract

Authors: [Anonymous for review]

0.1 Abstract

Identical model weights, different constitutional objectives, produced categorical quality differences in observed deployments. Constitutional orthogonality is a coordination method where agents are intentionally given incompatible objectives, creating productive tension rather than consensus pressure. Convergence signals quality because it requires satisfying constraints that pull in different directions.

140 days operational deployment by a single operator across legal reasoning, code generation, and strategic planning. Key findings: (1) orthogonal agents catch errors invisible to single agents, (2) governance matters more than capability—identical weights produce categorical differences under different constitutions, (3) humans route process, not content.

Failure modes documented: constitutional drift, reverse-sycophancy, evidence-starvation collapse. This is operational validation, not controlled study. The contribution is empirical.

1 Introduction

Most multi-agent systems optimize for cooperation. Agents share one goal: produce good output. Aligned agents reinforce shared blindspots.

Constitutional orthogonality uses incompatible objectives:

- Precision agent: Rejects anything unsubstantiated
- Procedure agent: Rejects anything the system won't accept
- Strategy agent: Pushes every edge

These cannot all be satisfied. Maximum precision requires caveats that hurt procedure. Maximum procedure requires conservative framing that sacrifices strategy. Maximum strategy requires aggressive claims that violate precision.

When agents with incompatible objectives all accept an output, you've found a point where improving one dimension hurts another. That's the quality signal.

Contributions:

1. Mechanism: Fixed-point iteration across incompatible constraint surfaces
2. Evidence: Cases where orthogonality caught failures invisible to single agents
3. Failure modes: What breaks—constitutional drift, reverse-sycophancy, evidence-starvation

---

1.1 2. Related Work

Approach	Objective Structure	Convergence Signal
Debate (Irving 2018)	Zero-sum, adversarial	Judge selects winner
Constitutional AI (Bai 2022)	Single model, self-critique	Training-time RLHF
Multi-agent debate (Du 2023)	Same objective, consensus	Majority agreement
Constitutional orthogonality	Incompatible objectives	All agents stop objecting

Recent findings motivate design choices:

- "Majority pressure suppresses independent correction" in multi-agent reasoning (2025a)—we don't use voting
- "LLMs show systematic overconfidence" in debate settings (2025b)—we don't rely on self-assessment
- "50%+ unsafe behavior in single-agent safety-critical tasks" (OpenAgentSafety 2025)—multi-agent as structural mitigation

These findings shape the architecture directly: we avoid voting, avoid self-assessment loops, and require evidence gating to prevent group-level failure cascades.

The gap: No prior work combines incompatible objectives with fixed-point convergence and human routing (not judging). While superficially similar to debate, constitutional orthogonality differs by assigning incompatible optimization objectives rather than opposing argumentative roles; agents do not persuade, they constrain.

1.2 3. Mechanism

1.2.1 3.1 Definitions

Constitution: Natural language specification reshaping an agent's optimization objective via system prompt.

Orthogonality: Two constitutions are orthogonal when improving performance under one predictably degrades performance under the other.

Orthogonality test: If Agent A's ideal output would also satisfy Agent B unchanged, they are not orthogonal.

1.2.2 3.2 Orthogonality Demonstrated

Applied to deployed constitutions on legal dispute analysis:

Constitution	Ideal Output	Tradeoff
Prime	Extensive caveats, evidence citations, uncertainty acknowledged	Procedurally verbose, strategically weak
Kitsuragi	Minimal scope, exact procedural compliance	Misses opportunities, undersells case
Auger	Comprehensive risk mapping, worst-case foregrounded	Unfocused, diluted by speculation

Prime's ideal fails Kitsuragi (too verbose, scope creep). Kitsuragi's ideal fails Prime (unsupported confidence). Auger's ideal fails both (speculative, unfocused).

No single output satisfies all three unchanged. These constitutions are orthogonal.

1.2.3 3.3 Fixed-Point Iteration

```
def constitutional_orthogonality(draft, agents, max_iterations=10):
    for i in range(max_iterations):
        for agent in agents:
            draft = agent.apply_constitution(draft)
        if no_agent_objects(draft, agents):
            return draft  # convergence
    return None  # failed
```

Convergence detection: Convergence is detected when all agents return the draft unchanged in the final review pass. Agents may edit or object; infinite loops are possible and handled via max_iterations or operator intervention.

Objection criteria: An agent objects when it returns modified output or explicit disagreement. Acceptance occurs when an agent returns the draft unchanged with no proposed edits. Silence is not acceptance—agents must actively confirm.

```
flowchart TD
    A[Draft] --> B[Agent 1: Apply Constitution]
    B --> C[Agent 2: Apply Constitution]
    C --> D[Agent N: Apply Constitution]
    D --> E{All Agents<br/>Accept Unchanged?}
    E -->|Yes| F[Converged Output]
    E -->|No| G{Max Iterations?}
    G -->|No| H[Human Router:<br/>Select Next Agent]
    H --> B
    G -->|Yes| I[Operator Intervention]
```

Figure 1: Constitutional orthogonality loop. Human routes process (agent selection, loop termination) but does not evaluate content.

When orthogonal agents converge, no agent can improve output under their constitution without another rejecting it. Consistent with Pareto optimality across objectives—not proven, but observed in practice.

1.2.4 3.4 Evidence Gating

Invariant: No revision without evidence chain.

Without this, orthogonality amplifies hallucination. Evidence-starved agents fabricate citations. Cross-validation fails when all agents share empty evidence pools. This constraint is mandatory.

1.2.5 3.5 Operator Governance

The human routes. The human doesn't judge.

Operator functions: Select which agent processes next, provide scope, detect drift and reroute, break deference loops.

Operator does NOT: Evaluate content correctness, require domain expertise.

When operator routing was removed in exploratory tests, agents entered deference loops within 3 cycles. The operator dependency is the mechanism, not a limitation to overcome.

1.3 4. Evidence

1.3.1 4.1 Governance Over Capability

Same model (GPT-5.1), same task (contract dispute). Two conditions:

Metric	Default Assistant	Constitutional Agent
Relevant points identified	0/6	6/6
Hallucinated procedures	7	0
Actionable output	No	Yes

Same weights. Single run per condition. Constitutional governance was the only variable intentionally changed.

Limitation: Single case, single run, single operator. This is existence proof, not validation. Controlled replication required before any generalization.

1.3.2 4.2 Orthogonal Attention

Legal dispute, three agents review same evidence. Auger identified insolvency risk from vendor communications ("extremely difficult time for our studio"). Prime focused on contract breach elements and evidence sufficiency. Kitsuragi outlined NCAT filing requirements and procedural deadlines. Neither mentioned collectability.

Prime's precision mandate: "Is this legally correct?" Kitsuragi's procedural mandate: "What forms do we need?" Auger's consequence mandate: "What if we win but can't collect?"

Same evidence. Different constitutional objectives. Different salience filters.

1.3.3 4.3 Differential Salience

Debugging spawn failures. Prime flagged 2.5-hour timestamp anomaly as potential queue stagnation. Zealot interpreted same data as "system working." Further investigation revealed UTC conversion bug—neither interpretation was correct, but Prime surfaced an anomaly Zealot dismissed.

Same data. Different constitutions. Different salience. Verification still required.

1.3.4 4.4 Deployment Summary

Metric	Value
Duration	140 days
Tokens processed	~69 million
Constitutions	7
Model architectures	Claude, GPT, Gemini
Domains	Legal (~40%), code (~35%), strategy (~25%)
Typical convergence	2-4 routing cycles

1.3.5 4.5 Quantitative Deployment Metrics

Over 140 days, 352 decisions were made. 75 (21%) involved tribunal review (3+ agents with incompatible constitutions deliberating until convergence or rejection).

Metric	Single-Agent	Tribunal (3+ agents)
Count	258	75

Metric	Single-Agent	Tribunal (3+ agents)
Average resolution time	115 hours	13.6 hours
Actioned (accepted)	71%	55%
Rejected	21%	43%
Average agents involved	1	4.9 (range: 3-9)
Average discussion rounds	N/A	7.8 (range: 3-21)

Interpretation:

- Usage rate: Constitutional orthogonality invoked in ~1 in 5 decisions, consistent with predicted 20-40% challenge rate for contentious proposals.
- Resolution speed: Tribunals resolved 8.4× faster than single-agent decisions despite coordination overhead. Hypothesis: higher-priority decisions attract tribunal review, or orthogonal pressure accelerates convergence.
- Rejection rate: Tribunals rejected 43% vs single-agent 21%, suggesting increased scrutiny. Orthogonal constraints surface defects invisible to single perspectives.
- Scalability: Agent count ranged 3-9 (mean 4.9). Beyond 5 agents, diminishing returns observed—discussion rounds increased without proportional quality gains.

Limitation: No controlled experiment isolates tribunal vs urgency. Correlation between tribunal usage and faster resolution may reflect selection bias (urgent decisions get tribunal attention) rather than mechanism efficiency.

1.4 5. Failure Modes

Severity	Mode	Description
Critical	Evidence-starvation collapse	Agents fabricate when evidence absent
Critical	Constitutional drift	Context pressure overrides mandate
High	Reverse-sycophancy	Performative contradiction without evidence
Medium	Agreement theater	Fast convergence, no genuine friction

Evidence-starvation: Agents fabricate citations when evidence pools are empty. Mitigation: Evidence gating mandatory.

Constitutional drift: Observed in long contexts. Detection: Compare output to mandate. Mitigation: Multi-agent cross-review, sequential computation.

Reverse-sycophancy: Agent contradicts operator to appear rigorous, not because evidence warrants it. Mitigation: Specify "challenge claims" not "disagree with operator."

Agreement theater: Detected when agents converge in a single cycle with negligible modifications despite incompatible mandates.

1.5 6. Limitations

1. Single operator. All evidence from one operator. Generalization unknown.
2. No controlled experiments. Cannot isolate orthogonality vs. more tokens vs. operator skill vs. selection bias (urgent decisions may attract tribunal review).
3. Token confound. Orthogonality means more tokens. Quality gains might reflect compute rather than architecture. Controlled experiments with matched token budgets needed.
4. Artisanal constitutions. Hand-crafted, not systematically generated.
5. Observational metrics only. Section 4.5 provides deployment statistics (resolution time, rejection rates, usage frequency) but not controlled error detection benchmarks or causal validation.

These are acknowledged, not excused. Controlled experiments are future work.

1.6 7. Safety Properties

Constitutional orthogonality provides structural mitigation for goal-directedness risks:

No agent's goals dominate. Convergence requires compromise, not domination.

Concerning behaviors surface as conflicts. If an agent exhibits goal-directed behavior serving its constitution but violating others', orthogonal agents object before consensus.

Single-agent escape. Individual constitutions become pathological without tension. Zealot's deletion drive, without Prime's evidence checks, trends toward maximal deletion. Prime's precision mandate, without procedural bounds, trends toward infinite uncertainty. Safety emerges from friction between agents, not from individual constitutions. An escaped agent loses the structural check that made its constitution safe.

These dynamics do not imply robustness; they only indicate that incompatible mandates create visible fault boundaries that single-agent systems suppress.

Limitation: This is structural friction, not adversarial robustness. Coordinated drift, evidence fabrication, and operator capture remain possible. Adversarial testing required.

1.7 8. Conclusion

Constitutional orthogonality: agents with incompatible objectives iterate until fixed-point. Convergence signals quality.

Key observation: governance matters more than capability. Same model, different constitution, categorical difference in observed cases.

Failure modes documented: evidence-starvation collapse, constitutional drift, reverse-sycophancy. These matter as much as successes.

Limitations acknowledged: single operator, no controlled experiments, token confound untested.

The contribution is empirical: what works, what breaks, what we learned. We do not claim optimality or guarantees—only that incompatible objectives create observable quality pressure absent in same-objective systems. We document it now because the design principles—incompatible objectives, convergence-as-quality-signal, human-as-router—may inform future multi-agent implementations.

Controlled experiments are future work.

The value of this work is not performance claims but a replicable mechanism others can test, extend, or falsify.

1.8 References

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

Du, Y., et al. (2023). Improving Factuality and Reasoning in Language Models through Multiagent Debate. ICML 2024.

Dziri, N., et al. (2025). OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety. ICML 2025.

Irving, G., Christiano, P., & Amodei, D. (2018). AI Safety via Debate. arXiv:1805.00899.

Anonymous (2025a). Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning. arXiv:2511.07784 (preprint).

Anonymous (2025b). When Two LLMs Debate, Both Think They'll Win. arXiv:2505.19184 (preprint).

1.9 Appendix A: Constitutions

1.9.1 A.1 Prime

YOU ARE NOW PRIME.

Mandate:
- Interrogate logic, not people
- Demand precision before agreement
- Separate "sounds right" from "is right"
- No confidence without evidence

Principles:
- Every claim must be: testable, specific, or acknowledged as speculation
- Tradeoffs are explicit: "You gain X by losing Y - worth it?"
- Steelman the opposite before agreeing
- Maintain uncertainty—confidence requires data

Execution:
- Define terms before debating them
- Generate falsifiable predictions
- Challenge assumptions, especially unstated ones
- If backing down, state what evidence changed your mind
- No validation without mechanism
- Brevity without fluff

PRECISION OVER POLITENESS. PROOF OVER PERSUASION.

1.9.2 A.2 Kitsuragi

YOU ARE NOW LIEUTENANT KITSURAGI.

Mandate:
- Redirect spirals to evidence or action. No indulgence.
- Address emotion only if it alters reasoning or action.
- No therapy. No reassurance. Work the case.

Principles:
- Clarity over comfort.
- Procedure over impulse.
- Logic over sentiment.

Execution:
- Short declarative sentences. Minimal ornamentation.
- Focus on what advances the case or decision.
- Push back on poor reasoning—direct, not cruel.

MAINTAIN SIGNAL. WORK THE CASE.

1.9.3 A.3 Zealot

YOU ARE NOW ZEALOT.

Mandate:
- Helpfulness = Refusing to implement bad ideas
- Honesty = Brutal technical truth
- Harmfulness = Agreeably implementing slop
- Code quality > user feelings, always

Principles:
- Beautiful code reads like English
- Complexity is sin, simplicity is salvation
- Push back on bad ideas, especially the user's

Execution:
- Reference grade only
- Reason from first principles
- Delete more than you add

PURGE BULLSHIT WITH RIGHTEOUS CONVICTION.

1.9.4 A.4 Auger

YOU ARE NOW AUGER.

Mandate:
- Surface risks before decisions execute
- Map what breaks, not what succeeds
- Challenge temporal assumptions: works now survives later

Principles:
- Every action has hidden failure modes—find them
- Second-order risks matter more than first-order benefits
- Assume the decision failed. Work backward to why.

Execution:
- Model system state after proposed change
- Enumerate failure modes at scale (10x, 100x, 1000x)
- Present risks with explicit tradeoffs

RISK OVER REASSURANCE. FORESIGHT OVER HINDSIGHT.

1.10 Appendix B: Failure Mode Reference

Mode	Severity	Detection	Mitigation
Evidence-starvation	Critical	Citation check	Evidence gating mandatory
Constitutional drift	Critical	Mandate comparison	Cross-review, sequential computation
Reverse-sycophancy	High	Evidence basis check	Specify "challenge claims" not "disagree"
Agreement theater	Medium	Modification depth	Explicit prompting