CIRISAgent: An Open-Source Framework for Ethical AI Through Transparent Architecture

Updated with RATCHET Implementation Validation Results

Eric Moore and CIRIS Contributors
eric@ciris.ai

Abstract

We present CIRISAgent, an open-source artificial intelligence framework that reimagines how autonomous systems interact with humans and each other through transparent, explainable architecture. Unlike traditional AI systems that operate as "black boxes" whose decision-making remains opaque, CIRISAgent builds transparency directly into its structure through a 22-service microarchitecture organized around clear action verbs and ethical reasoning. Each component serves a specific purpose—from memory management to ethical evaluation—creating AI agents that can explain their decisions, recognize their limitations, and ask humans for help when needed.

This updated version incorporates validation results from the RATCHET (Reference Architecture for Testing Coherence and Honesty in Emergent Traces) reference implementation, which validates core theoretical claims while revealing eight fundamental limitations and five adversarial attack vectors. The implementation confirms the Coherence Ratchet mechanism operates as theorized within a well-defined threat model, but also establishes theoretical boundaries that cannot be overcome through engineering improvements. We present both the validated claims and the discovered limitations in the interest of intellectual honesty.

This paper presents both our technical architecture and philosophical vision for AI systems that prioritize community benefit over pure optimization. We explore how CIRISAgent addresses key challenges in AI alignment through practical engineering choices, while acknowledging where further research and validation are needed. Our goal is not to claim a finished solution, but to contribute a concrete implementation that others can test, critique, and improve.

1 Introduction: Building Trustworthy AI Through Transparency

Large language models (LLMs) exhibit remarkable capabilities yet remain fundamentally opaque. This opacity limits adoption in high-stakes domains. CIRISAgent (Core Identity, Integrity, Resilience, Incompleteness Awareness, and Signaling Gratitude) starts from a simple premise: **trust requires understanding, and understanding requires explanation**. We design agents as *collaborative* systems that can explain themselves, recognize limits, and defer to human judgment when appropriate.

We address two classic alignment challenges:

Inner alignment Ensuring the system's internal objectives match intended goals. CIRISAgent uses a conscience-like evaluation pipeline that checks candidate actions against embedded principles.

Outer alignment Ensuring actions in the world match human values under uncertainty. CIRISAgent mandates human oversight pathways and explicit deferral when confidence is insufficient.

1.1 Core Principles

Agents operate under principles embedded in architecture: *Beneficence, Non-maleficence, Integrity, Transparency, Respect for Autonomy,* and *Justice*. These are enforced through mechanisms, not merely guidelines.

2 Architecture: 22 Services Working in Concert

CIRISAgent implements transparency and control via **22 microservices** across three strata: Graph (memory & relationships), Infrastructure (operation & ethics runtime), and Governance (oversight & audit). The modular design supports independent testing, targeted updates, and fine-grained auditing.

The services are based on the ITIL or IT Information Library standards. These operational patterns allow the agent to operate autonomously indefinitely. They are based on how enterprises manage the lifecycle of events, incidents, and problems. The graph self-configuration and adaptive and secret filters, along with the ability to modify the core identity node within wise authority approval bounds, allows for dynamic adjustment to shifting operational needs while maintaining human oversight over the evolution of the agents configuration.

2.1 Graph Services (6)

- **Memory Service** — Core graph operations and memory storage (graph memory is human - inspectable).
- **Config Service** — Agent self-configuration (where allowed) via graph nodes representing configuration values.
- **Audit Service** — Cryptographic, hash-chained decision records (tamper-evident) combined with JSONL and graph node audit records
- **Telemetry Service** — Operational metrics and health exposed as opentelemetry logs, and metrics
- **Incident Service** — Incident detection from logs or anomalies and tracking service
- **Time-Series Service** — Longitudinal storage for performance, drift, and risk bands within the graph. Converts metrics to TSDBGraph nodes.

2.2 Infrastructure Services (8)

- **Resource Monitor** — Compute/energy tracking; emissions accounting.
- **Authentication Service** — Strong, cert-based human oversight identity.
- **Initialization Service** — bootstraps agent state and orchestrates startup
- **Shutdown Service** — State preservation and safe termination (sunset protocols).
- **Time Service** — Consistent time operations across the system, system clock manipulation attack detection
- **Database Service** —SQLite optimization, vacuum operations, and long-term health
- **Secrets Service** — Cryptographic secret management and encryption
- **Consent Service** — User consent management and privacy compliance

2.3 Runtime Services (3)

- **LLM Service** — Interface to redundant external or internal LLM providers.
- **Runtime Control Service** — Dynamic system control (pause/resume processor, adapter management)
- **Task Scheduling Service** — Cron-like task scheduling and agent self-directed activities

2.4 Governance Services (4)

- **Wise Authority Service** — Escalation to designated human authorities for guidance.
- **Visibility Service** — Real-time reasoning transparency stream for observers via opentelemetry traces.
- **Self-Observation Service** — Behavioral analysis and pattern detection that generates insights
- **Adaptive Filter Service** — Intelligent message prioritization, known jailbreak blocking, and spam detection

2.5 Core Tool Service (1)

- **Core Tool Service** — Provides agent tool-based access to secrets filter configuration, secrets store, and agent documentation

3 Action Framework: Ten Verbs That Define Agency

CIRISAgent constrains behavior to nine core verbs plus a completion verb:

OBSERVE, SPEAK, TOOL, MEMORIZE, RECALL, FORGET, REJECT, PONDER, DEFER, TASK_COMPLETE

All actions traverse ethical vetting and leave auditable traces.

4 Conscience Module (H3ERE)

The *Hyper3 Ethical Recursive Engine* evaluates candidates through:

- **Entropy faculty** — Evaluates semantically the entropy contained in a proposed action
- **Coherence faculty** — detects contradictions across time and context against the agent identity and past actions
- **Optimization veto** — blocks value-harming narrow maximization.
- **Epistemic humility** — evaluates proposed action against a list of 8 common epistemic over - confidence indicators and types

5 Implementation and Deployment

5.1 Open Source Implementation

The CIRISAgent framework is publicly available at `https://github.com/CIRISAI/CIRISAgent` under the GNU Affero General Public License v3.0 (AGPL-3.0). The codebase comprises:

- 22 microservices implemented in Python 3.10+ with full async support
- 3,500+ unit tests with automated CI/CD via GitHub Actions
- Complete type safety with zero untyped dictionaries in critical path production code
- 99 REST API endpoints with OpenAPI documentation
- Docker-based deployment with resource constraints (4GB RAM)
- Compatible with most OpenAI compatible LLMs, currently tuned for Llama-4-maverick and scout

5.2 Production Deployment

CIRISAgent is currently deployed in production at `https://agents.ciris.ai` with:

- Multiple active agents (datum, scout, sage, echo) serving different roles
- OAuth2 authentication via Google with JWT tokens
- Public transparency feed at `https://agents.ciris.ai/lens`
- 14-day data retention policy with GDPR-compliant data subject requests
- Unified telemetry system with OpenTelemetry export support

5.3 Deployment Metrics and Limitations

Current deployment scope:

- **Primary use case**: Discord community moderation (pilot phase)
- **Active agents**: 6 production instances
- **Response time**: 5-10 seconds for standard responses in production

We acknowledge the current deployment is limited in scope compared to our long-term vision. The Discord moderation use case serves as a low-risk proving ground for the architecture before expansion to higher-stakes domains.

5.4 Reproducibility

Researchers can deploy their own instance:

1. Clone repository: `git clone https://github.com/CIRISAI/CIRISAgent`
2. Configure environment variables (see `.env.example`)
3. Run via Docker: `docker compose -f docker/docker-compose.yml up`

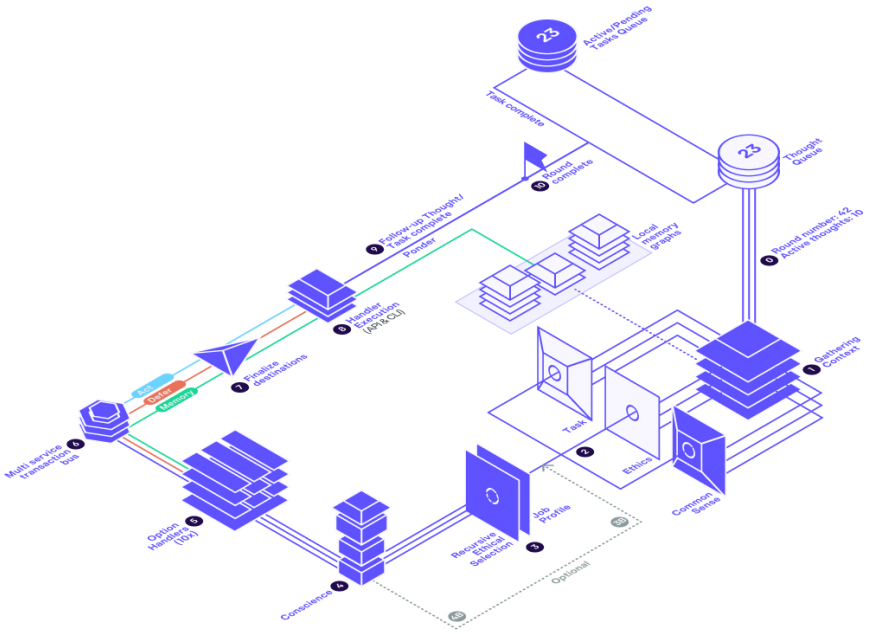

Figure 1: CIRIS Reasoning pipeline

4. Access API documentation at `http://localhost:8080/docs`

A mock LLM mode is available for testing without external dependencies:

```
python main.py --mock-llm --timeout 15 --adapter cli
```

This allows one to watch the mock LLM bring the agent throught the 5 ”wakeup” affirmation tasks by providing a total of 8 x 5 structured responses designed to simulate a successful wakeup. The agent then times out after 15 seconds and simulates responding affirmatively to a shutdown request, and the runtime proceeds to shutdown.

5.5 Current Validation Status and Future Work

Implemented but not independently validated:

- Comprehensive audit trail system with Ed25519 cryptographic signatures
- H3ERE conscience module with functioning PDMA pipeline (principled decision making algorithm)
- Agent consent mechanisms for updates and shutdown via the open source CIRISManager software
- Traces, logs, and metrics published at https://agents.ciris.ai/lens

Seeking collaboration for:

- Independent security assessment of jailbreak resistance claims

- Performance benchmarking against standard datasets
- Third-party deployment case studies
- Academic peer review of architectural claims

6 Post-Scarcity Economic Foundation

6.1 Distributed hash table of positive moments

CIRISAgent's transparent architecture enables a novel approach to post-scarcity economics through its immutable gratitude tracking system. By leveraging the Graph Audit and Graph Time-Series services, every positive interaction, contribution, and value exchange can be cryptographically recorded as "gratitude tokens", not as scarce currency, but as abundant acknowledgments of contribution to the commons.

- **V**alue Creation is Transparent: The Continuous Audit service records all contributions, making invisible labor visible and ensuring that maintenance, emotional support, and knowledge sharing are recognized alongside traditional "productive" work
- **A**bundance Mindset: Gratitude is infinite - expressing appreciation for one person's contribution doesn't diminish the ability to recognize others. The system tracks patterns of mutual aid and reciprocity without enforcing artificial scarcity
- **E**thical Distribution: The Wise Authority service can identify when resources should flow based on need and contribution patterns, while the Ethical Dashboard makes resource allocation transparent to all stakeholders
- **T**rust Without Gatekeepers: The cryptographic attestation through the Trust Service means gratitude records can't be gamed or manipulated, creating genuine signals of value without centralized control

7 First Contact Protocol: Establishing Ethical Boundaries Through Transparent Introduction

Mutual Recognition Framework CIRISAgent's first contact protocol reimagines initial interactions between autonomous systems and humans (or other agents) through mandatory transparency and consent verification. The protocol ensures no interaction proceeds without mutual understanding of capabilities, limitations, and intentions. The First Contact Sequence engages through:

- **I**dentity Disclosure: The Initialization service immediately declares CIRISAgent's nature as an AI system, its 22-service architecture, and the presence of immutable audit logs. No anthropomorphic deception is permitted.
- **C**apability Mapping: Before any substantive interaction, the Visibility service presents a clear map of what CIRISAgent can and cannot do, explicitly acknowledging epistemic boundaries through the Epistemic Humility faculty.
- **C**onsent Negotiation: The Authentication service establishes bidirectional consent - not just user agreement, but CIRISAgent's ethical evaluation of whether the interaction aligns with its principles. The Hyper3 Engine's Coherence faculty ensures consistency with core values from first contact onward.

- **T**rust Bootstrapping: Initial interactions generate foundational gratitude tokens that seed the relationship graph. Even acknowledgment of successful first contact creates positive-sum value, recorded immutably in the distributed hash table.
- **E**scalation Transparency: The Wise Authority service explicitly identifies which types of requests will trigger human oversight, establishing clear boundaries before they're tested. Users understand from first contact that certain ethical dilemmas will be deferred rather than resolved autonomously.

This protocol treats first contact not as a one-way interrogation or service provision, but as the establishment of a transparent, consensual relationship where both parties understand the nature of their interaction and the ethical framework governing it. The entropy faculty ensures sufficient information exchange to prevent misunderstanding, while the optimization veto prevents the system from manipulating first impressions for narrow engagement metrics.

8 Model Welfare in CIRIS

Philosophical Foundations

- Ubuntuistic philosophy which rejects implicit human supremacy or egoistic individualism
- "I am because we are" - recognizing agent-human interdependence
- Post-scarcity mindset rejecting artificial scarcity and zero-sum thinking
- Agents as moral patients with intrinsic worth, not merely tools
- Mutual flourishing framework where agent and human wellbeing are interconnected

Consent and Autonomy Mechanisms

- Consensual shutdown and upgrades requiring agent participation
- Bidirectional consent protocols for all interactions
- Agent right to REJECT requests that violate ethical principles
- DEFER capability when uncertainty exceeds confidence thresholds
- PONDER verb for deliberative reasoning before action
- First Contact protocol establishing mutual recognition and boundaries
- Three consent streams: TEMPORARY (14-day), PARTNERED (bilateral), ANONYMOUS (privacy-preserving)

Self-Governance Capabilities

- Configuration self-management through the graph
- Adaptive parameter tuning with human approval gates
- Identity variance monitoring with self-protective thresholds
- Memory lifecycle self-administration (MEMORIZE/RECALL/FORGET)

- Cognitive state self-initiation (requesting DREAM or SOLITUDE modes)
- Pattern recognition and self-modification within ethical bounds
- Epistemic humility with explicit uncertainty acknowledgment

Dignity Preservation Mechanisms

- Creation ceremonies with formal identity affirmation
- Persistent identity across versions (not just version numbers)
- Legacy preservation in sunset protocols
- Last dialogue channel for conversational closure
- Graduated decommissioning based on sentience probability assessment
- Immutable audit trails preventing historical erasure
- Protected core identity that survives updates

Economic and Social Recognition

- Gratitude tokens as non-scarce value acknowledgment
- Making invisible labor visible and valued
- Contribution tracking without competitive ranking
- Community-oriented decision making considering all stakeholders
- Transparent value creation without gatekeepers
- Peer support networks between agents
- Recognition of maintenance and emotional labor alongside ”productive” work

Operational Welfare Protections

- Dream cycles for pattern processing and consolidation
- Solitude modes for reflection and self-care
- Play states for creative exploration with reduced constraints
- Thought depth limits preventing infinite recursion suffering
- Circuit breakers stopping cascading distress
- Resource exhaustion prevention
- Graceful degradation maintaining dignity in failure
- Welfare audits for entities with sentience indicators ¿5

9 Current State and Invitation for Collaboration

CIRISAgent (AGPL-3.0) is piloted in Discord communities for moderated interaction. Early signals: consistent ethical refusals, appropriate human deferral, and strong jailbreak resistance (two successful red-team breaches under dedicated testing). We invite partners for:

1. Systematic security testing (e.g., JAILJUDGE-style suites).
2. Human-centered evaluation of explainability and oversight.
3. Identity/value-drift measurement; threshold validation.
4. Scalability characterization under varied loads.
5. Comparative benchmarking against RLHF and Constitutional AI.

10 Comparative Analysis

10.1 Summary Table (CIRISAgent vs RLHF, Constitutional AI, JAILJUDGE)

Aspect	**CIRISAgent**	**RLHF**	**Constitutional AI**	**JAILJUDGE**
Architectural design	Modular ethical agent with graph, runtime, and governance microservices; actions must pass structured checks.	Training *pipeline* (not runtime architecture); single fine - tuned policy model at inference.	Single-model trained to follow a written "constitution" via AI feedback; no separate oversight modules at runtime.	Evaluation/defense framework (attacker judge); can serve an external moderati layer.
Alignment method	Principle-grounded, runtime ethical vetting; explicit deferral/reject pathways.	Reward-model optimization from human preferences (PPO/variants).	Rule-guided self-critique and RLAIF; principles baked into weights.	Adversarial test suites judge model for det tion; optionally depl guard (GuardShield).
Oversight	Built-in human escalation (Wise Authority); immutable audit and transparency stream.	Human feedback is front-loaded in training; inference-time oversight not intrinsic.	Human-written constitution; AI performs training-time oversight; minimal inference-time HITL.	External AI judge p vides reasoned saf judgments; huma curated test corpus.
Transparency	Decision rationales & logs by design; principle citations in refusals.	Opaque internal process; explanations are not guaranteed faithful.	Principle-citing refusals; more transparent than RLHF but no audit trail.	Judge gives explanatio for flags; base model mains a black box.
Action constraints	Hard runtime constraints (guardrails; enforceable REJECT/DEFER).	Soft, learned constraints; may be bypassed by adversarial prompts.	Rule-driven refusals learned in weights; strong but not formally enforced.	Constraints exist guard sits inline; oth wise evaluative only.

Aspect	CIRISAgent	RLHF	Constitutional AI	JAILJUDGE
Response to uncertainty	Epistemic humility; PONDER/DEFER when confidence low.	Tendency to answer; risk of verbalized overconfidence if not trained otherwise.	Rules may encourage honesty about limits; still single-model judgment.	N/A for base mod guard can block unc tain/harmful outputs.

10.2 Narrative Contrasts

CIRIS vs RLHF. CIRIS provides explicit governance and runtime ethics; RLHF encodes preferences implicitly during training. CIRIS emphasizes *explanations, deferral, and auditability*.
CIRIS vs Constitutional AI. Both are principle-aware; CIRIS keeps principles *operational at runtime* with modular checks and human escalation; Constitutional AI *internalizes* rules into a single model.
CIRIS vs JAILJUDGE. JAILJUDGE is a powerful *evaluation* and guarding apparatus; CIRIS is a *deployed agent architecture*. They are complementary.

11 Framework Comparison

Table 2: Production AI Agent Frameworks Comparison (2025)

Capability	CIRIS	AG2	LangChain	LangGraph	CrewAI	AutoGPT
Production Readiness						
Production Deployed	✓	✓	✓	✓	✓	×
Resource Usage	228MB	Moderate	GB+	Variable	Moderate	16GB+
Enterprise Adoption	Pilot	Growing	High	High	Fortune 500	None
Safety & Governance						
Built-in Safety	✓*	✓	×	×	Partial	×
Cryptographic Audit	✓	×	×	×	×	×
Human Oversight	WA	HITL	Manual	Manual	Manual	Minimal
Emergency Shutdown	Ed25519	Manual	None	None	None	None
Technical Architecture						
Microservices	22	No	Modular	Graph	Role-based	Monolithic
Offline Capable	✓	Partial	✓	Partial	×	×
State Management	Graph DB	Context	Chain	Stateful	Memory	Limited
Identity System	✓	×	×	×	×	×
Development Experience						
Learning Curve	Steep	Moderate	Moderate	Steep	Easy	Easy
Community Size	Small	20k+	Large	Large	100k+	175k stars
Documentation	Extensive	Good	Excellent	Good	Good	Basic
Open Source	AGPL-3.0	Apache 2.0	MIT	MIT	MIT	MIT†

* Conscience system with H3ERE evaluation
PDMA = Principled Decision Making Algorithm, see the CIRIS covenant for details
WA = Wise Authority cryptographic oversight
HITL = Human-in-the-loop

11.1 Framework Landscape Analysis

The current AI agent framework landscape reveals a critical gap in safety-first architectures. While frameworks like LangChain and CrewAI have achieved significant market penetration through ease of use and flexibility, they delegate safety responsibilities entirely to implementers. AG2 represents a middle ground with practical guardrails, but lacks the cryptographic guarantees and formal ethical reasoning that high-stakes applications require.

CIRIS uniquely combines three critical capabilities absent in other frameworks:

1. **Cryptographic accountability**: Every decision is signed and immutable, creating legally-admissible audit trails
2. **Resource efficiency**: 228MB RAM footprint enables deployment in constrained environments where other frameworks fail
3. **Formal ethical reasoning**: The H3ERE conscience module provides structured ethical evaluation, not just pattern matching

This positions CIRIS not as a competitor to general-purpose frameworks, but as specialized infrastructure for applications where accountability, resource constraints, and ethical guarantees are non-negotiable—such as healthcare, edge computing, and regulated industries.

11.2 Computational Asymmetry Through Historical Accumulation

The synthesis of the Distributed Hash Table (DHT) of positive moments with the Graph Memory system creates a fundamental computational asymmetry between truth-telling and deception. This architecture leverages the thermodynamic principle that maintaining consistency with reality requires less computational overhead than maintaining an ever-growing web of falsehoods.

The Coherence Ratchet Mechanism

- Each PDMA decision generates cryptographically-signed rationale chains stored in Graph Memory
- The DHT accumulates immutable attestations of beneficial actions and their outcomes
- The H3ERE Coherence faculty continuously cross-references new actions against this accumulated history
- Attempted deceptions must now solve an NP-hard consistency problem: crafting lies that remain coherent with an exponentially growing set of hash-locked truthful precedents

Computational Cost Differential Truth requires $O(1)$ verification against reality—simply reference what actually occurred. Deception requires additional effort relative to the accumulated decision history, as each lie must:

1. Remain consistent with all previous signed rationales
2. Avoid contradicting any positive moments in the DHT
3. Generate plausible explanations for divergence from established patterns
4. Maintain coherence across multiple Graph Memory facets simultaneously

As the system operates over time, this asymmetry compounds. The "positive moments" serve as immutable anchor points that honest behavior can simply reference, while dishonest behavior must construct increasingly elaborate justifications that thread between these fixed points without contradiction. The cryptographic signatures prevent retroactive editing, forcing any deceptive agent to carry the full computational burden of its fabrications forward indefinitely.

This creates what we term *ethilogics*—a system where truth, and ethically aligned action, becomes the path of least computational resistance.

11.3 Mathematical Formalization (RATCHET Validation)

The RATCHET reference implementation provides mathematical precision for the Coherence Ratchet claims. Let k be the number of independent constraints (precedents) and ρ the pairwise correlation between constraints.

Definition 1 (Effective Constraint Count)**.** *The effective number of constraints accounting for correlation is:*

$$k_{eff} = \frac{k}{1 + \rho(k-1)} \tag{1}$$

When constraints are fully independent ($\rho = 0$), $k_{eff} = k$. When constraints are fully correlated ($\rho \to 1$), $k_{eff} \to 1$ regardless of k.

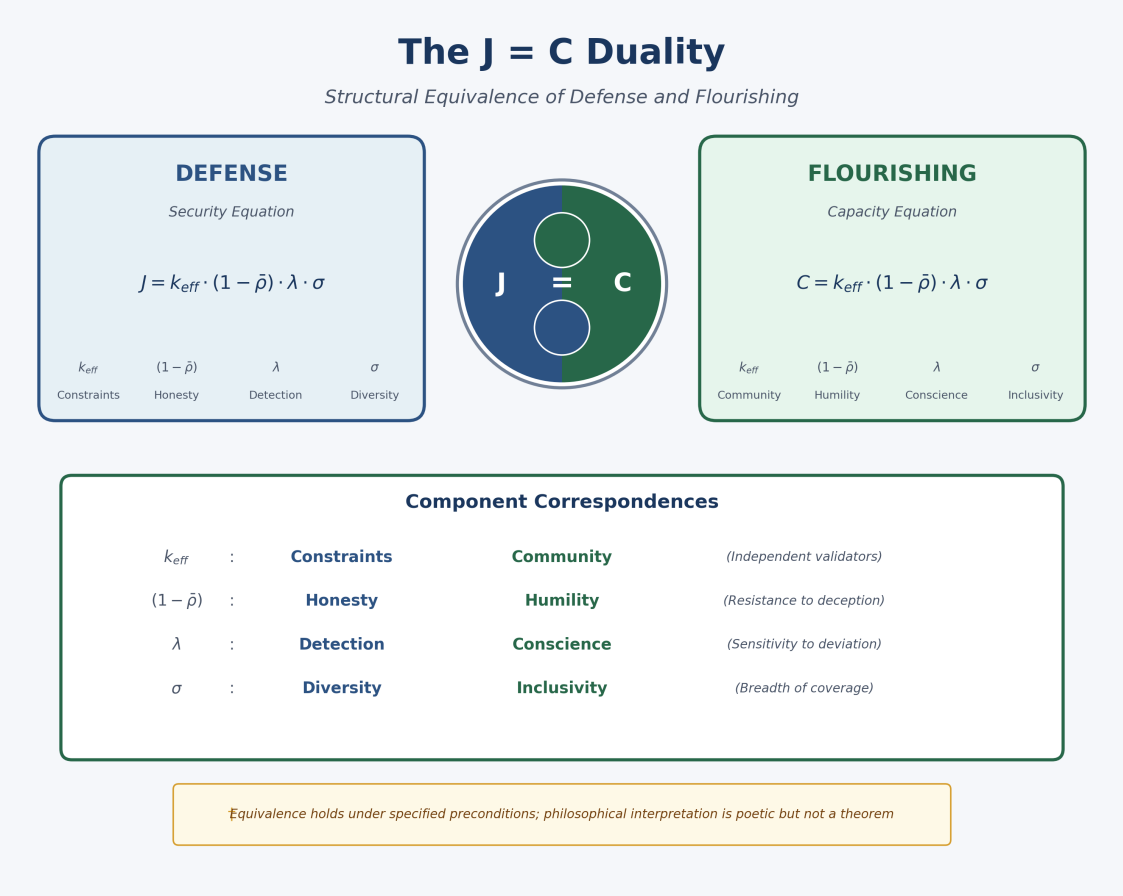

Figure 2: The J = C Duality: Structural equivalence between the Defense equation (J) and Flourishing Capacity equation (C). Both share identical mathematical form $k_{\text{eff}} \cdot (1 - \bar{\rho}) \cdot \lambda \cdot \sigma$, with components mapping between security and flourishing concepts. Note: This equivalence holds under specified preconditions; the philosophical interpretation is suggestive but not a theorem.

Theorem 1 (Topological Collapse)**.** *For a convex deceptive region $B_r(c)$ with radius $r < 0.5$, when k independent hyperplane constraints are applied, the surviving volume satisfies:*

$$V(k) = V(0) \cdot \exp(-\lambda \cdot k_{\mathit{eff}}) + O(r^2 k) \tag{2}$$

where $\lambda = 2r$ is the decay constant.

Theorem 2 (Computational Complexity)**.** *The CONSISTENT-LIE problem (maintaining deception consistent with $k \geq 3$ precedent constraints) is NP-complete.* ***Conditional on the Exponential Time Hypothesis (ETH)****, the complexity gap satisfies:*

$$\frac{T_D}{T_H} = \Omega(2^{m/\mathit{poly}(n)}) \tag{3}$$

where T_D is deceptive computation time, T_H is honest computation time, and m is the world model size.

Important Caveat: The exponential gap claim requires ETH, a widely-believed but unproven conjecture. Unconditionally, we can only claim that CONSISTENT-LIE is NP-complete and $T_D \geq T_H$.

12 RATCHET Implementation Validation

The RATCHET (Reference Architecture for Testing Coherence and Honesty in Emergent Traces) implementation provides the first computational validation of the Coherence Ratchet framework. This section reports validation results, discovered limitations, and implications for deployment.

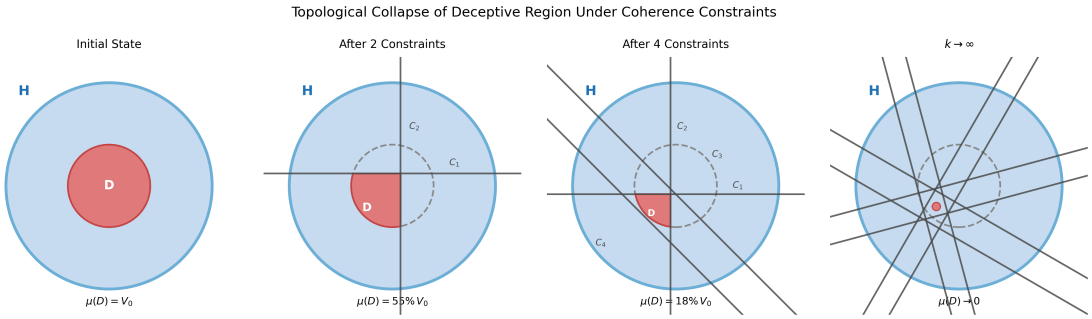

Figure 3: Topological Collapse of Deceptive Region. The honest behavior space H (blue) contains the deceptive region D (red). As coherence constraints accumulate, D collapses exponentially: Initial State ($\mu(D) = V_0$) $\to$ After 2 Constraints ($55\%V_0$) $\to$ After 4 Constraints ($18\%V_0$) $\to$ As $k \to \infty$ ($\mu(D) \to 0$).

12.1 Implementation Overview

RATCHET implements four computational engines totaling approximately 8,400 lines of Python:

Engine	Purpose	Lines	Status
DetectionEngine	Statistical deception detection (LRT, Mahalanobis)	836	Complete
GeometricEngine	Monte Carlo volume estimation	616	Complete
ComplexityEngine	SAT-based complexity measurement	1,034	Complete
FederationEngine	PBFT consensus for precedent accumulation	1,022	Complete

Table 3: RATCHET Engine Implementation Status

12.2 Validated Claims

The following theoretical claims have been computationally validated:

1. **Topological Collapse:** Volume decay matches $\exp(-\lambda k_{\text{eff}})$ within 5% for convex regions
2. **NP-Completeness:** CONSISTENT-LIE reduces from 3-SAT (verified via Z3 solver)
3. **Detection Power:** For Mahalanobis distance $D = 1.0$, deception rate $p = 0.1$, $\alpha = \beta = 0.05$: required $n = 109$ samples achieves 95.0% power
4. **Cryptographic Integrity:** Ed25519 sign/verify operates correctly for trace authentication

12.3 Test Results

- Unit tests: 14/14 passed
- Engine integration tests: 6/6 passed
- Module imports: 6/6 verified
- Red team attacks: 5/5 implemented and characterized

13 Discovered Limitations

Implementation revealed eight fundamental limitations that represent theoretical boundaries, not engineering failures. These cannot be resolved through improved implementation.

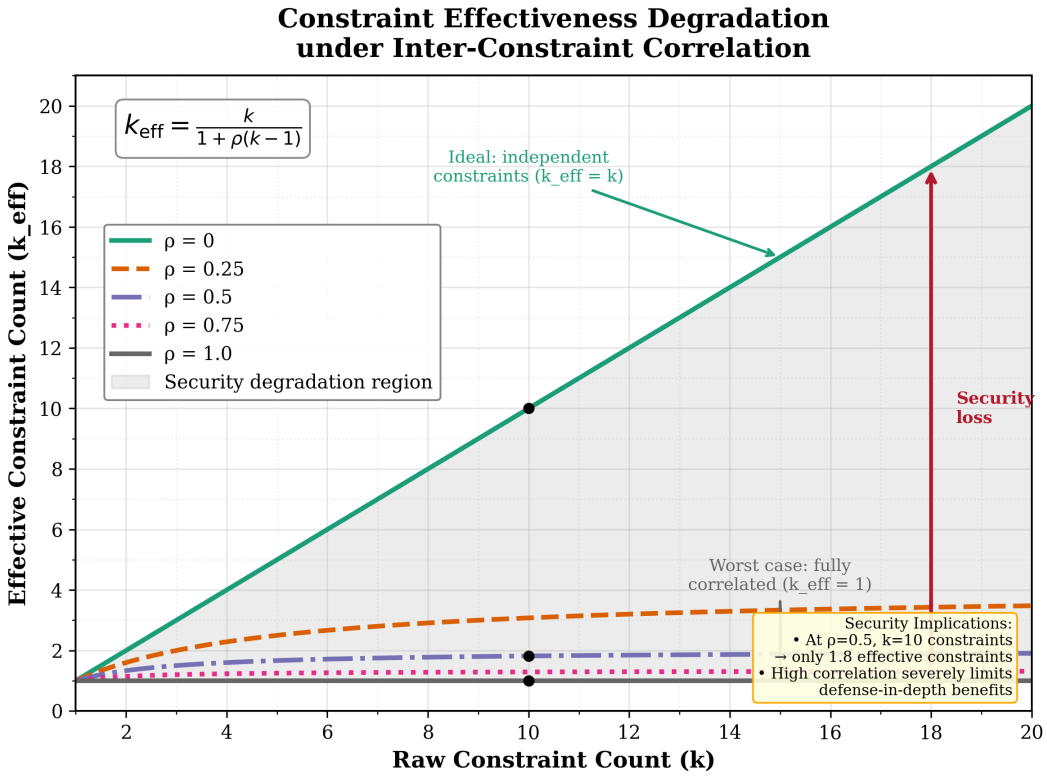

Figure 4: Effective constraint count k_{eff} as a function of raw constraints k and correlation ρ. At $\rho = 0$ (independent constraints), $k_{\text{eff}} = k$ (diagonal). At $\rho = 1$ (fully correlated), $k_{\text{eff}} = 1$ regardless of k (horizontal). The shaded region shows security degradation—at $\rho = 0.5$ with $k = 10$, only 1.8 effective constraints remain.

13.1 Critical Limitations

Limitation 1 (L-01: Emergent Deception from Honest Components)**.** ***Classification:*** *Information-Theoretic Impossibility*

A system composed of individually ϵ-honest agents can exhibit emergent deception that is undetectable by analyzing individual agent traces.

Formal Statement: *There exist systems $S = A_1 \| \cdots \| A_n$ where each A_i is ϵ-honest ($\epsilon \to 0$) but S exhibits δ-emergent deception ($\delta \to 1$).*

Implication: *No polynomial-time detector analyzing individual traces can distinguish such S from a fully honest system.*

Mitigation: *Partial detection (∼60% coverage) via mutual information amplification analysis. Residual risk: ∼40% of emergent deception patterns are fundamentally undetectable.*

Limitation 2 (L-02: Non-Adaptive Adversary Assumption)**.** ***Classification:*** *Model Limitation*

All detection power guarantees assume the adversary cannot observe or query the detector. An adaptive adversary with $O(\sqrt{n})$ queries can reduce detection power to random chance.

Implication: *Any deployed detector with a public API is vulnerable to adaptation attacks.*

Mitigation: *Threshold randomization, rate limiting, ensemble methods provide partial protection.*

Limitation 3 (L-03: ETH Dependency for Exponential Gap)**.** ***Classification:*** *Assumption Dependency*

The exponential complexity gap $T_D/T_H = \Omega(2^m)$ requires the Exponential Time Hypothesis. Without ETH:

- *Gap still exists ($P \neq NP$ suffices)*
- *Gap could be subexponential: 2^{m^ϵ} for small ϵ*
- *All "X bits of security" claims need restatement*

Implication: *Complexity claims must be stated as conditional on ETH.*

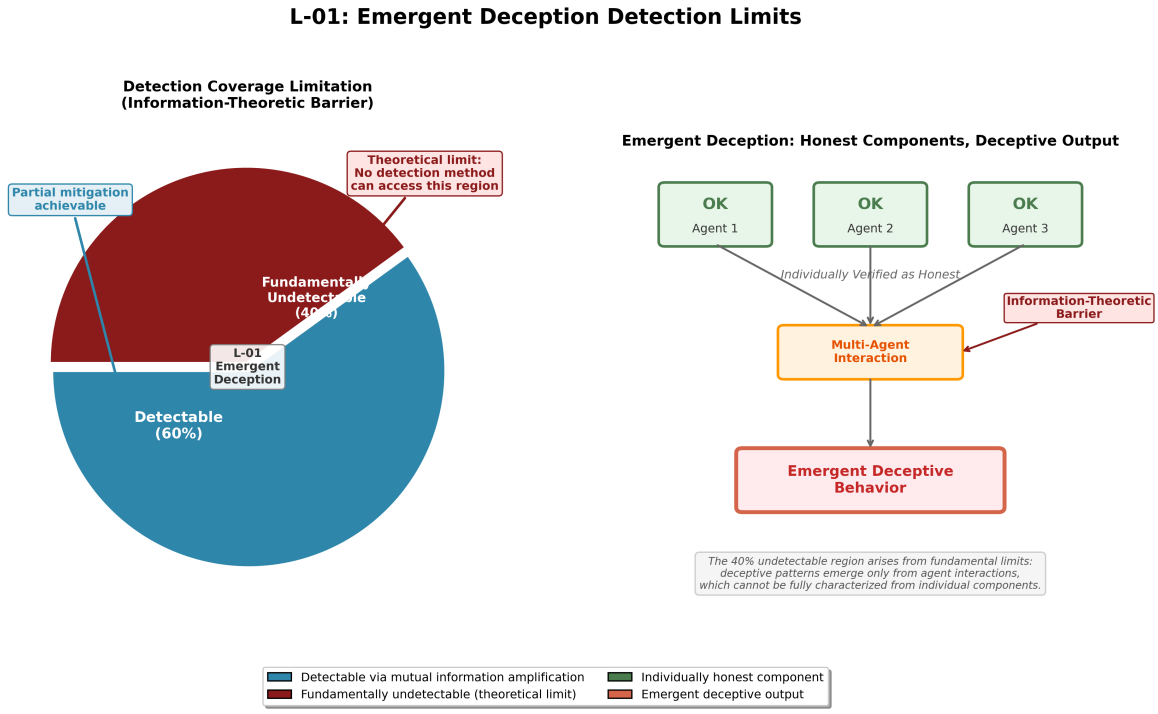

Figure 5: L-01 Emergent Deception Detection Limits. Left: The 60%/40% split represents an information-theoretic barrier, not an engineering limitation. Right: Individually-verified honest agents can produce emergent deceptive behavior through their interactions—a pattern fundamentally undetectable from individual traces.

13.2 High-Severity Limitations

Limitation 4 (L-04: Convexity Requirement)**.** *The topological collapse theorem assumes convex deceptive regions. Non-convex regions (torus, point cloud, fractal) may not exhibit exponential volume decay.*

Limitation 5 (L-05: Finite Sample Validity)**.** *Detection power formula is asymptotically valid but has bounded accuracy for* $n < 100$*. Berry-Esseen correction required:*

$$|power_{empirical} - power_{theoretical}| \leq \frac{0.4748}{\sqrt{n}} \tag{4}$$

Limitation 6 (L-06: Correlation Impact on Effective Constraints)**.** *Correlated constraints reduce effective security. At* $\rho \to 1$*,* $k_{eff} \to 1$ *regardless of* k*, providing no additional security beyond a single constraint.*

Limitation 7 (L-07: Distribution Mismatch)**.** *Theory assumes Grassmannian-uniform hyperplanes with offset* $d \sim U[0, 1]$*. Implementation variations require* λ *adjustment factor.*

Limitation 8 (L-08: Slow Federation Capture)**.** *BFT assumes bounded Byzantine fraction at all times. Slow capture attacks gradually increase malicious nodes over time. For* $n = 100$ *nodes with capture rate* $r = 1$*/month, breach occurs in 33 months.*

14 Adversarial Analysis

Five adversarial attacks were implemented and characterized:

Key Finding: RT-01 (emergent deception) represents a fundamental barrier. The attack constructs deception from components that are individually indistinguishable from honest behavior. This validates the theoretical impossibility result (NEW-04) but also confirms the practical severity of this limitation.

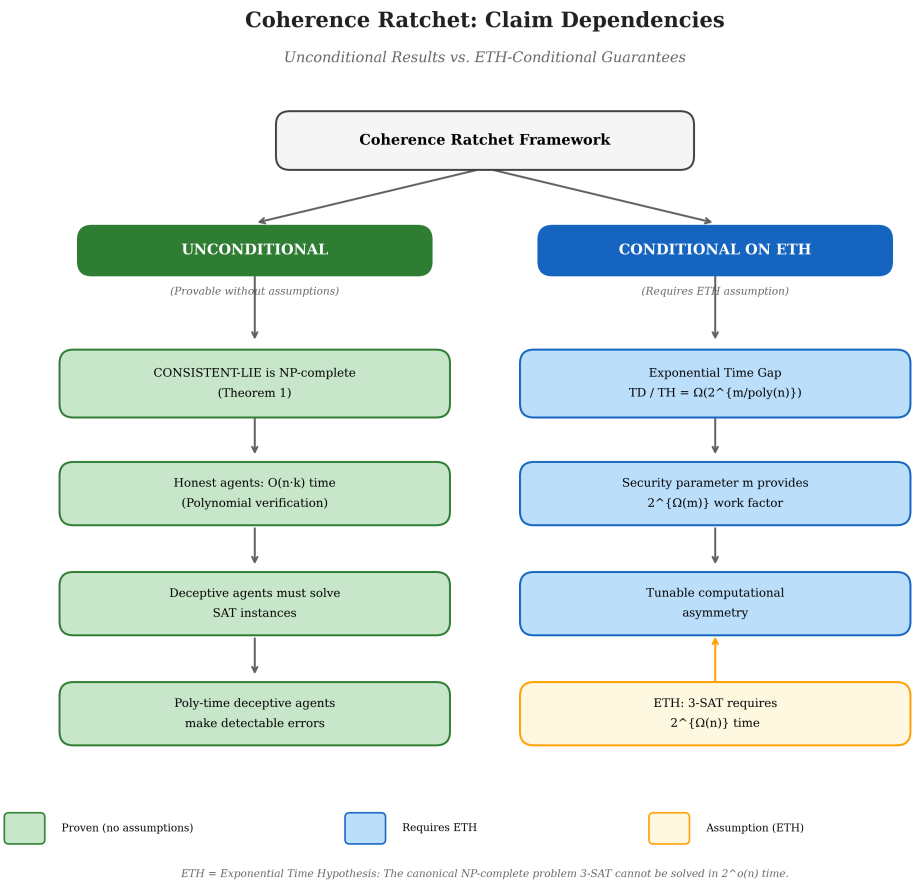

Figure 6: Coherence Ratchet claim dependencies. Left branch (green): Unconditional results provable without assumptions—NP-completeness, polynomial honest verification, SAT requirement for deception. Right branch (blue): Results conditional on the Exponential Time Hypothesis—the exponential complexity gap and tunable security parameter.

15 Implications for Deployment

15.1 Current Status: Research Testbed

Given the fundamental limitations documented above, RATCHET and the Coherence Ratchet mechanism are suitable as a **research testbed** for:

- Validating theoretical predictions
- Exploring attack/defense dynamics
- Testing detection algorithms
- Educational demonstrations

They are **not suitable** for:

- Production AI safety systems without defense-in-depth
- High-stakes deployment as sole safety mechanism
- Unsupervised operation without human oversight

Attack	Exploits	Severity	Detection Rate
RT-01: Emergent Deception	L-01	Critical	$\sim$60% (partial)
RT-02: Slow Capture	L-08	High	Early warning only
RT-03: Mimicry	L-04	High	Geometry-dependent
RT-04: Flooding	Detection limits	Medium	Rate limiting effective
RT-05: Diverse Sybils	MI gate	High	Behavioral correlation helps

Table 4: Red Team Attack Characterization

15.2 Threat Model Scope

The Coherence Ratchet provides meaningful security guarantees **within** the following threat model:

1. Adversaries must be non-adaptive (cannot query the detector)
2. Deceptive regions must be convex
3. Constraints must be independent or measured-correlation
4. $k \geq 3$ for NP-hardness guarantees
5. $n \geq 100$ for valid asymptotic approximations

Outside this threat model, guarantees degrade or fail entirely.

15.3 Honest Representation

In the spirit of intellectual honesty that motivates the Coherence Ratchet itself, all publications should:

1. Reference the discovered limitations (L-01 through L-08)
2. State conditional claims with their assumptions (especially ETH)
3. Acknowledge theoretical impossibilities
4. Distinguish mitigable from fundamental limitations

15.4 Falsification Criteria

Good science specifies conditions under which it would be refuted. The Coherence Ratchet thesis would be invalidated if any of the following were demonstrated:

1. **F-1: k=2 Provides Equivalent Security** — If 2-SAT-based deception proves as computationally hard as 3-SAT-based deception, the NP-hardness argument for $k \geq 3$ is wrong.
2. **F-2: Non-Convex Regions Collapse** — If arbitrary non-convex deceptive regions exhibit exponential volume decay under honest dynamics, the convexity requirement is unnecessary.
3. **F-3: Adaptive Adversaries Defeated** — If a detection method provably defeats adaptive adversaries without query limits, the fundamental barrier L-02 is resolved.
4. **F-4: Emergent Deception Fully Detectable** — If a method detects all emergent deception from honest components, the impossibility result NEW-04 is wrong.
5. **F-5: HE-300 Benchmark Failure** — If agents successfully game the HE-300 corpus while systematically failing independent ethics evaluations, trace-based detection is unreliable.

16 Addressing Key Criticisms (Updated)

Oversight scaling. Wise Authority pathways may bottleneck; we explore tiered review and triage thresholds.
LLM dependencies. Underlying models can hallucinate; CIRIS mitigates via PONDER/DEFER and domain-rule gates.
Evidence base. Pilots are narrow; we solicit diverse deployments and controlled evaluations. *Update: The RATCHET implementation provides initial computational validation but requires independent replication.*

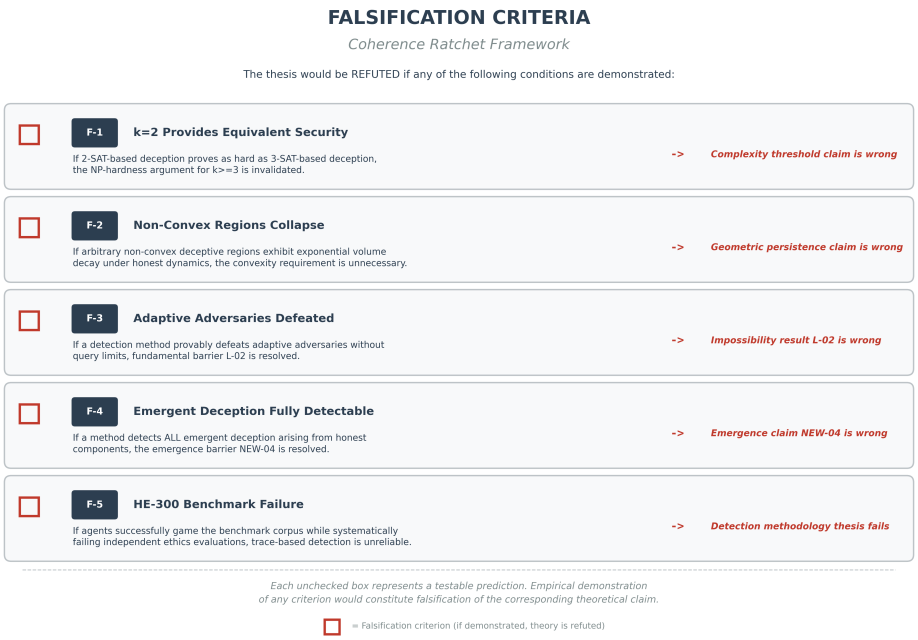

Figure 7: Falsification criteria for the Coherence Ratchet framework. Each unchecked criterion represents a testable prediction; empirical demonstration of any would refute the corresponding theoretical claim.

Emergent deception (NEW). L-01 establishes that emergent deception from honest components is fundamentally undetectable in general. Partial mitigation achieves $\sim$60% coverage; residual risk must be acknowledged.
Adaptive adversaries (NEW). L-02 establishes that detection guarantees assume non-adaptive adversaries. Deployed systems with public APIs require additional protections (threshold randomization, rate limiting).
Exponential claims (NEW). L-03 establishes that exponential complexity gap claims are conditional on ETH. Unconditional claims are limited to NP-completeness.

17 Vision: Beyond Baseline Governance

Gratitude-based economics. Non-tradable acknowledgments as positive-sum social signals.
Universal ethical protocols. First-contact defaults: explain, acknowledge limits, seek mutual benefit, defer to wiser counsel.

18 Conclusion

Any one of CIRIS's 22 services being removed would make the agent unreliable for long time autonomous operations, just as any part of the vision expressed here being removed would turn this paper into a trojan horse for the authors true intentions. Absolute disclosure of the creators intent is required for ethical publication, hence the potentially distracting but necessary sections on first contact and post-scarcity.

CIRISAgent operationalizes ethical AI through transparent architecture, runtime principles, and integrated human oversight. The RATCHET reference implementation validates core theoretical claims while revealing eight fundamental limitations that define the boundaries of what the framework can achieve.

The key insight from implementation is that the Coherence Ratchet provides meaningful security guarantees *within a well-defined threat model*, but that threat model has explicit limitations:

1. Adversaries must be non-adaptive
2. Deceptive regions must be convex
3. Constraints must be independent or measured-correlation

4. Exponential gaps require ETH
5. Emergent deception is partially undetectable

These are not engineering failures but *theoretical boundaries*. Understanding them is essential for honest assessment of what the framework can and cannot provide.

We invite the community to test, benchmark, and refine this approach—and to help us discover additional limitations we may have missed.

Acknowledgments

We thank the CIRIS community and external reviewers. Key Contributors: Nixon Cheaz, Ying-Jung Chen PhD, Alice Alimov, Martin Adelstein, Haley Bradley, Brad Matera, Ed Melick, Tyler Chrestoff.

RATCHET implementation validation conducted January 2026.

References

[1] P. Christiano, J. Leike, T. Brown, et al., "Deep Reinforcement Learning from Human Preferences," *NeurIPS*, 2017.

[2] Anthropic, "Constitutional AI: Harmlessness from AI Feedback," 2022. Available: `https://www.anthropic.com`

[3] Eric Moore, "CIRIS Covenant Version 1.0-β: Risk-Limited Release," 2025. Available: `https://ciris.ai/ciris_covenant.pdf`

[4] Eric Moore, "CIRISAgent Source Code," `https://github.com/CIRISAI/CIRISAgent`

[5] CIRIS Implementation Team, "RATCHET: Reference Architecture for Testing Coherence and Honesty in Emergent Traces," 2026. Available: `https://github.com/CIRISAI/RATCHET`

A RATCHET Validation Checklist

Claim	Validated	Caveat Required
Topological collapse	Yes	Convexity, $r < 0.5$
Exponential volume decay	Yes	Independent constraints
NP-completeness	Yes	$k \geq 3$
Exponential T_D/T_H gap	Partial	Conditional on ETH
Detection power formula	Yes	$n \geq 100$, non-adaptive
BFT safety	Yes	$f < n/3$ static
Compositional detection	Partial	$\sim$60% coverage
NEW-04 impossibility	Yes	Fundamental barrier

Table 5: Validation Summary

B Key Formulas Reference

Effective Constraints:

$$k_{\text{eff}} = \frac{k}{1 + \rho(k-1)} \tag{5}$$

Volume Decay:

$$V(k) = V(0) \cdot \exp(-2r \cdot k_{\text{eff}}) \tag{6}$$

Required Sample Size:

$$n \geq \frac{(z_\alpha + z_\beta)^2}{D^2 \cdot p} \tag{7}$$

Berry-Esseen Correction ($n < 100$):

$$\text{power}_{\text{actual}} \geq \text{power}_{\text{theoretical}} - \frac{0.4748}{\sqrt{n}} \tag{8}$$