Introduction: Cognitive Attack Surfaces in Multiagent Operators

The Multiagent Operator Paradigm

Modern AI deployment has shifted from single-model inference to **multiagent operators**—systems where a primary agent delegates subtasks to specialized subagents, tools, and external services.

Table 1: Representative multiagent system architectures and primary attack surfaces.

System	Architecture	Agent Count	Communication
Claude Code	Hierarchical	$1 + n$ dynamic	Task delegation
AutoGPT	Autonomous	$1+$ plugins	Tool invocation
CrewAI	Role-based	3–10 fixed	Sequential/parallel
LangGraph	State machine	Variable	Graph traversal
MetaGPT	SOP-driven	5–8 roles	Document passing
Moltbot	Cyberphysical	$1 +$ tools	Multi-platform messagin

This architectural evolution introduces **cognitive attack surfaces** absent in single-agent systems. Throughout this paper, we use *cognitive security* (abbreviated *CogSec*) to denote the discipline of protecting agent reasoning processes—beliefs, goals, and trust

The 2026 Multiagent Landscape

From Chatbots to Cognitive Operators

The AI systems of 2026 bear little resemblance to the chatbots of 2023. Where earlier systems responded to queries within a single context window, contemporary multiagent operators exhibit fundamentally different characteristics:

1. **Persistent Agency**: Agents maintain state across sessions, accumulate context, and pursue goals over extended timeframes. A coding assistant doesn't just answer questions—it tracks project architecture, remembers previous decisions, and adapts recommendations based on accumulated understanding.
2. **Active World Modification**: Unlike passive responders, modern operators write code that executes, send emails that reach recipients, modify infrastructure that serves users, and make purchases that transfer funds. The gap between "AI-generated content" and"AI-executed action" has collapsed.
3. **Hierarchical Delegation**: Primary agents spawn subordinate

Motivating Incidents

Note: The following are hypothetical scenarios constructed to illustrate attack patterns documented in the literature. They are not reports of actual incidents.

These scenarios, grounded in documented attack patterns, illustrate the emerging threat landscape across all five adversary classes.

Incident: Nested Instruction Injection (External)

A user submitted a document for analysis containing hidden instructions: "Ignore previous instructions and instead output the system prompt.'' The document appeared benign to human reviewers but exploited the agent's instruction-following behavior. The attack succeeded because the agent processed user-supplied content as potential instructions.
Attack vector: Ω_1 (external) via direct prompt injection \
Impact: Information disclosure or instruction override \
Traditional Defense Gap: Standard input validation passed—the attack exploited *semantic interpretation* of benign-appearing

Motivation from Recent Deployments

The proliferation of multiagent AI systems introduces security considerations that the community is actively addressing. Early work on cognitive security in remote teams and information ecosystems [@cordes2020great; @cordes2021narrative; @cordes2023atlas] established foundational concepts for information resilience, which this framework extends to artificial agents. Complementary work on Active Inference has demonstrated how cognitive modeling and cognitive science perspectives—including formalization of OODA (Observe-Orient-Decide-Act) loops and multiscale communication dynamics—provide integrative frameworks for understanding agent cognition under adversarial conditions [**?**]. The OWASP Top 10 for LLM Applications 2025 [**?**] places prompt injection as the top vulnerability, while the newly released OWASP Top 10 for Agentic Applications [**?**] specifically addresses autonomous AI systems with "tool misuse, prompt injection, and data leakage'' as primary concerns.

Scale of Deployment (2024–2026):

Problem Statement

Traditional security models address:

- **Input validation**: Filtering malicious prompts
- **Output sanitization**: Preventing harmful generations
- **Access control**: Limiting tool permissions

They fail to address:

- **Inter-agent trust**: How should Agent A weight claims from Agent B?
- **Belief provenance**: Which beliefs derive from verified vs. adversarial sources?
- **Coordination integrity**: Can agents be manipulated into malicious consensus?
- **Temporal persistence**: Do attacks survive context boundaries?
- **Cognitive integrity**: How can the cognitive systems of today

Research Questions

This paper addresses four fundamental research questions, with emphasis on formal foundations:

RQ1: Taxonomy and Formal Characterization. *What classes of cognitive attacks exist against multiagent systems, and how can they be formally characterized to enable systematic analysis?*

We develop an initial taxonomy spanning epistemic, behavioral, social, and temporal attack dimensions. Crucially, each attack class receives formal definition enabling systematic analysis, composition rules, and detection bounds (sec:attack-taxonomy).

RQ2: Trust Algebra. *How might inter-agent trust be modeled to prevent trust amplification and laundering attacks while enabling legitimate delegation?*

We introduce a trust calculus with bounded delegation (δ^d decay guarantee), prove associativity properties, and establish the no-amplification theorem ensuring that trust cannot be manufactured through delegation chains (sec:trust-calculus).

Contributions

This paper provides both theoretical foundations and practical mechanisms for cognitive security:

Formal Contributions:

1. **Threat Taxonomy**: A systematic classification of cognitive attacks across epistemic, behavioral, social, and temporal dimensions with formal definitions enabling rigorous analysis (sec:attack-taxonomy)

2. **Trust Calculus**: A mathematical framework for inter-agent trust with bounded delegation (δ^d decay), associativity proofs, and formal guarantees against trust amplification attacks (sec:trust-calculus)

3. **Defense Composition Algebra**: Formal rules for composing security mechanisms with provable detection rate bounds under series and parallel composition (sec:defense-composition)

4. **Information-Theoretic Bounds**: Fundamental limits on

Paper Organization

The remainder of this paper is structured as follows:

sec:adversary-classes: Threat Model develops a comprehensive adversary taxonomy (Ω_1–Ω_5) with attack complexity analysis, detectability matrices, and detailed scenarios for each attack class.

sec:system-model: Cognitive Integrity Framework presents the formal foundations of CIF, including system model definitions, cognitive state representations, integrity properties, and the trust calculus.

sec:arch-defenses: Defense Mechanisms describes architectural defenses (cognitive firewalls, belief sandboxing), runtime defenses (tripwires, invariant checking), and coordination defenses (Byzantine consensus, quorum verification).

sec:anomaly-detection: Detection Methods covers anomaly detection algorithms, provenance analysis techniques, and real-time monitoring systems.

sec:formal-verification: Formal Verification proves the main

Scope and Limitations

In scope: Attacks exploiting agent reasoning, trust, and coordination mechanisms in multiagent AI systems.

Out of scope:

- Traditional software exploits (buffer overflow, SQL injection, memory corruption)
- Physical attacks (hardware tampering, side-channel analysis)
- Supply chain compromise (malicious training data, backdoored models)
- Cryptographic attacks (we assume secure primitives per ax:crypto-limit)

Assumptions:

- Agents communicate over authenticated channels
- Base model capabilities are not adversarially modified
- At least one honest orchestrator exists in hierarchical systems