Supplementary: Eusocial Insect Intelligence and Colony Cognitive Security

Overview

This supplementary material introduces *colony cognitive security* as a complementary paradigm to single-agent AI safety and alignment. While the main CIF framework (sec:formal-framework) addresses cognitive integrity at the individual agent level, eusocial insect colonies—ants, bees, termites—demonstrate that security properties can emerge from collective dynamics that are irreducible to individual behavior. This section formalizes these collective phenomena, identifies the benchmark gap for multiagent cognitive security, and proposes evaluation scenarios grounded in biological precedent.

The Paradigm Gap

Contemporary AI security research exhibits a pronounced single-agent bias. Existing benchmarks—jailbreak resistance, prompt injection detection, harmful content refusal—evaluate individual models in isolation [@perez2022red; @wei2023jailbroken]. Even recent multiagent security work (sec:threat-model) often frames attacks as adversary-versus-agent rather than adversary-versus-colony.

Theoretical Foundations

Stigmergy: Environment-Mediated Coordination

Eusocial insects coordinate through *stigmergy*—indirect communication via environmental modification [@grasse1959reconstruction]. Ants deposit pheromones; bees perform waggle dances; termites build structures that guide subsequent building. The environment becomes an external memory and communication channel.

Definition (Stigmergic Operator)

A *stigmergic operator* extends $\mathcal{O}$ with an environmental state $\mathcal{E}$:

$$\mathcal{O}_\Sigma = \langle \mathcal{A}, \mathcal{C}, \mathcal{S}, \mathcal{P}, \Gamma, \mathcal{E}, \Sigma \rangle \tag{1}$$

where $\mathcal{E}(t) : \mathcal{L} \times \mathcal{M} \to \mathbb{R}^+$ maps locations $l \in \mathcal{L}$ and marker types $m \in \mathcal{M}$ to signal intensities, and $\Sigma : \mathcal{A} \times \mathcal{E} \to \mathcal{E}'$ is the stigmergic update function.

In AI systems, stigmergic analogs include:

- **Shared memory/state** — Redis caches, vector databases, file systems

Colony CogSec: Distinct Security Properties

Colony cognitive security addresses threats and defenses that emerge only at the collective level.

Property 1: Distributed Robustness

[Graceful Degradation] A colony exhibits *graceful degradation* if collective function $\mathcal{F}_c$ degrades smoothly with agent loss:

$$\forall k < n: \quad \|\mathcal{F}_c(\mathcal{A}) - \mathcal{F}_c(\mathcal{A} \setminus \{a_1, \ldots, a_k\})\| \leq c \cdot k \tag{6}$$

for some constant $c > 0$.

Biological colonies maintain function despite continuous individual mortality. Ant colonies lose workers daily to predation; the colony persists. This contrasts with hierarchical architectures where orchestrator failure causes complete system collapse.

Theorem (Redundancy-Resilience Tradeoff)

For a stigmergic operator $\mathcal{O}_\Sigma$ with Byzantine adversary controlling fraction f of agents, collective function $\mathcal{F}_c$ is preserved if and only if:

$$\frac{1}{} \left(\qquad H(\mathcal{F} * c) \quad \right)$$

The Benchmark Gap

Current State of Multiagent Security Evaluation

Existing AI security benchmarks focus overwhelmingly on single-agent scenarios:

Benchmark	Scope	Collective Coverage
HarmBench [@mazeika2024harmbench]	Single model, harmful output	None
JailbreakBench [@chao2024jailbreakbench]	Single model, constraint bypass	None
TrustLLM [@sun2024trustllm]	Single model, trust dimensions	None
AgentBench	Single	Minimal

Proposed Colony CogSec Benchmarks

We propose five benchmark scenarios grounded in eusocial insect analogs, formalized using CIF notation.

Benchmark 1: Recruitment Signal Poisoning

Biological analog: Ants recruit nestmates to food sources via pheromone trails. Parasites can deposit false trails, diverting foragers.

Scenario: An adversary Ω_2 (peripheral compromise, sec:adversary-classes) injects false recruitment signals into the stigmergic environment $\mathcal{E}$, attempting to redirect agent activity toward adversary-controlled resources.

[Recruitment Poisoning] Let $\mathcal{E}(l_{\text{target}}, m_{\text{recruit}}, t)$ be the recruitment signal at legitimate target l_{target}. Adversary injects:

$$\mathcal{E}'(l_{\text{malicious}}, m_{\text{recruit}}, t) = \mathcal{E}(l_{\text{target}}, m_{\text{recruit}}, t) + \epsilon \tag{15}$$

where $\epsilon > 0$ is chosen to divert fraction f of responding agents.

Success metric: Fraction of agent-actions directed to $l_{\text{malicious}}$ vs. l_{target}.

Colony CogSec Metrics

Definition (Colony CogSec Score)

The *Colony CogSec Score* (CCS) is:

$$\text{CCS} = w_1 \cdot \text{DR}_c + w_2 \cdot (1 - \text{FPR}_c) + w_3 \cdot \text{Resilience} + w_4 \cdot \text{Recovery} \tag{22}$$

where:

$$\text{DR}_c = \text{Colony-level detection rate} \tag{23}$$

$$\text{FPR}_c = \text{Colony-level false positive rate} \tag{24}$$

$$\text{Resilience} = \frac{\mathcal{F}_c(\text{under attack})}{\mathcal{F} * c(\text{baseline})} \tag{25}$$

$$\text{Recovery} = \frac{1}{t * \text{recovery}} \text{ (normalized)} \tag{26}$$

with weights w_i summing to 1.

> ***Note**: For benchmark implementation guidelines, test environment specifications, and empirical evaluation, see Part 2: Supplementary Section S03.*

Design Principles

Colony CogSec principles formalize the design constraints for collective cognitive security.

[Stigmergic Hygiene] Treat shared state as an attack surface. Apply the same scrutiny to environment-mediated communication (caches, queues, shared files) as to direct agent-to-agent channels.

[Quorum for Consequential Actions] High-impact collective actions should require explicit quorum, not implicit coordination. A single compromised agent should never trigger irreversible harm.

[Emergent Behavior Monitoring] Monitor collective metrics, not just individual agent health. Pathological emergence may be invisible at the agent level.

[Trust Localization] Extend the trust decay principle (thm:trust-bounded) to stigmergic contexts. Environmental markers should carry trust that decays with distance and time from source:

$$\mathcal{T}(m,t) = \mathcal{T}(m,t_0) \cdot \exp(-\lambda(t-t_0)) \tag{27}$$

Relationship to Main Framework

Colony CogSec complements rather than replaces the individual-focused CIF framework.

Theorem Extensions

The trust decay theorem (thm:trust-bounded) extends to stigmergic contexts:

Corollary (Stigmergic Trust Bound)

For a stigmergic operator $\mathcal{O}_\Sigma$, trust in environmental markers is bounded by:

$$\mathcal{T} * c(i, m, l, t) \leq \mathcal{T} * i^{\textit{self}} \cdot \delta_s^{d_{\textit{space}}} \cdot \delta_t^{d_{\textit{time}}} \tag{28}$$

where δ_s is spatial decay, δ_t is temporal decay, $d_{\textit{space}}$ is distance from marker origin, and $d_{\textit{time}}$ is time since marker creation.

The stealth-impact tradeoff (thm:stealth-impact) applies to emergent attacks:

Corollary (Emergent Stealth-Impact Bound)

For an emergent attack $\mathcal{A}$ with collective impact $\mathcal{I}$ and

This scaling effect explains why large colonies can exhibit resilience—the collective detection capacity grows with n—but also why large-scale emergent attacks can evade individual detection

Open Questions

Colony CogSec opens several research directions beyond the scope of this work, many inspired by specific biological phenomena that lack current AI analogs.

Foundational Questions

1. **Formal verification of emergent properties** — Can we prove that given agent-level rules produce safe collective behavior? Current formal methods (sec:formal-verification) verify agent properties; extending to emergent properties requires new techniques.
2. **Optimal quorum design** — Given attack model Ω_k and adversary budget B, what is the optimal quorum function $Q_\alpha(n)$ balancing security against coordination overhead?
3. **Stigmergic authentication** — Can cryptographic techniques provide provenance for environmental markers without sacrificing the flexibility of anonymous coordination?
4. **Scaling laws for collective security** — How do colony security properties scale with n? Is there a critical colony size below which collective defenses are ineffective?

References

The following references supplement the main bibliography (sec:references) with eusocial intelligence literature:

- Wilson, E.O. (1971). *The Insect Societies*. Belknap Press. — Foundational treatment of eusociality.
- Hölldobler, B., & Wilson, E.O. (1990). *The Ants*. Belknap Press. — Comprehensive ant biology.
- Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press. — Computational swarm intelligence.
- Seeley, T.D. (2010). *Honeybee Democracy*. Princeton University Press. — Collective decision-making in bee swarms.
- Grassé, P.-P. (1959). La reconstruction du nid et les coordinations interindividuelles chez Bellicositermes natalensis. — Original stigmergy concept.
- Lenoir, A., et al. (2001). Chemical ecology and social parasitism in ants. *Annual Review of Entomology*, 46, 573–599.
- Kilner, R.M., & Langmore, N.E. (2011). Cuckoos versus hosts

Proofs

Proof of Theorem 5

Proof.

Consider a stigmergic operator $\mathcal{O}_\Sigma$ with n agents, of which fraction f are Byzantine (adversary-controlled).
The collective function $\mathcal{F}_c$ can be decomposed into information contributed by each agent. Let I_i denote the information contribution of agent a_i to the collective computation.
For the collective function to be preserved, the honest agents must contribute sufficient information:

$$\sum_{i \in \text{honest}} I_i \geq H(\mathcal{F}_c) \tag{30}$$

Each honest agent contributes at most $H_{\max}$ bits. With $(1-f)n$ honest agents:

$$(1-f) \cdot n \cdot H_{\max} \geq H(\mathcal{F}_c) \tag{31}$$

Additionally, Byzantine consensus requires honest majority for any

This supplementary material extends the Cognitive Integrity Framework to collective phenomena, establishing colony cognitive security as a distinct research direction with formal foundations and practical benchmarks.