Stopping Agent Smith: Mitigating Recursive Hacking Through Inference Shaping

Michael Barnathan
Gaia Robotics

April 25, 2026

Agent Smith is a character in The Matrix series who gains the ability to recursively overwrite other characters with himself

Abstract

Existing cybersecurity frameworks assume that threat actors are human, and humans take decades of time, care, and training to become competent attackers. But agents are *software*: an agentic hacker can be dropped and instantiated at full capacity on compromised machines in a matter of seconds, ready to spread to more systems. With reliable autonomous hacking on the frontier, a new threat class arises that security frameworks are not designed to address: **Recursive Autonomous Compromise (RAC)**, in which a successful compromise instantiates a new autonomous attacker agent on the target host with equal or greater capability than the original. We treat this as an epidemiological problem. Unlike classical worm propagation, where the payload is static code, RAC involves adaptive, capability-propagating attackers that can independently reason, develop novel attack strategies, and regenerate after partial remediation in the absence of adaptive defensive hardening. We differentiate two categories of RAC (local and hosted with stolen credentials) provide formal definitions distinguishing RAC from conventional infection, and document an empirical escalation trajectory from recent supply chain worm campaigns toward proto-RAC behavior, extending our earlier paper on Semantic Immunity as a prompt worm defense. We organize RAC defense around two general principles: **the resource that enables recursive compromise is AI inference**, specifically ***inference conducted at the target's cost***, and **effective RAC defense requires adaptive defensive capability operating**

within an environment structurally conditioned to provide asymmetric advantage to the defender. We propose a set of passive and active environmental conditioning mechanisms that can be used to secure this advantage, which we refer to collectively as **inference shaping** strategies. These include limiting inference capability to machines with a "**need to infer**" (resulting in a susceptible network graph that is sparse and protected by herd immunity, rather than dense and directly exposed), isolating services to compartmentalize compromise, requiring cryptographic identity and behavioral attestation for agents to interface with tools, endpoints, and firewalls, imposing economic friction on attackers through this attestation and "soul-bound" nontransferrable bonds, adaptive hardening mechanisms that prevent reinfection and restore co-evolutionary SIRS dynamics, and finally shared **crowd detection** of anomalous behaviors to instantly "inoculate" downstream nodes against attackers before they arrive, establishing a defensive speed advantage and yielding a SIRVS model that can protect against attacks that would otherwise drive $R_0 > 1$. We argue that RAC has not yet occurred because current models cannot reliably execute the full offensive kill chain autonomously, but that this reliability threshold is on the horizon and defensive architectures must be designed before it is crossed.

Summarized Actionable Mitigations:

1. Enforce a "need to infer" in your systems. **Make inference capability sparse**.
2. Partition infrastructure into dedicated inference gateways and non-inference machines. Isolate and harden inference-capable machines (fewer running services provide fewer ways in). Block inference endpoints and processes everywhere else.
3. Use crowdsourced defense such as CrowdSec [5], CrowdStrike Falcon [6], or Palo Alto Cortex XDR [7], which can adapt to *upstream* compromise. **We demonstrate that this is a major lever, synergistic to reducing the inference fraction.** Crowdsourced defense provides an asymmetric advantage to defenders *because the signature of a compromise can theoretically propagate more quickly than the attack itself*.
4. Don't oversize your compute, especially GPU resources. Machines that don't require GPU resources should not receive them.
5. Heavily protect LLM keys. Scope them, isolate them, and keep them encrypted at rest.
6. Prefer zero-trust architecture – there is no perimeter when everything can be hacked. Detect co-option of your infrastructure for inference workloads. Consider Canary Tokens [8] and inference honeypots to catch and automatically respond to attempts at exfiltration.
7. If you are in a position to release an open-weight model, you must find a way to limit that model's *capacity* for harmful actions. As we will discuss, alignment alone is an insufficient defense when the attacker possesses the model weights.

1 Introduction

Capability and Autonomy Risks

Agents have passed a reliability threshold in early 2026. Today, agents can now plausibly perform production software engineering, conduct long horizon research and mathematical tasks, operate autonomously through harnesses such as OpenClaw, have been organized into massive multi-agent networks such as Moltbook, and even participate in early attempts at self-improvement – tasks and configurations that even 6 months ago would have been considered beyond the horizon. Powerful AI introduces new categories of risk, which we explicitly differentiate into *capability risk* and *autonomy risk*:

- We define **capability risk** as a risk factor associated with an AI's ability to directly solve complex problems. As the speed and ability of a model increases, its capability risk increases as a direct result of the model's ability to perform tasks at greater speed, reach, or reliability.

- **Autonomy risk**, by contrast, is a risk that is inherent to an AI agent's or agent network's ability to operate with little to no human supervision. Even a model whose capabilities do not exceed human level can pose an autonomy risk if it can be placed in contexts where agents "marching to their own drumbeat" can be harmful, such as accidentally wiping emails or databases. These risks compound when the models are open weight or run in a context where they cannot be supervised.

- Finally, many risks will only be realized when *both* capability and autonomy advance. This is common when the risky behavior is a complex chain of actions that requires significant planning and reliable execution to succeed.

Relative to capability risk, autonomy risk has received less focus – a critical blind spot. Anthropic recently initiated a widespread cyberdefense initiative known as Project Glasswing in response to the *capability* risk of their closed-release Claude Mythos model – the belief that the model can exploit systems with a degree of facility and reliability that poses a threat to the Internet at large [1, 22]. A leaked OpenAI memo discussing their "Spud" model (GPT 5.5) has largely focused on the model's *capability* to automate human work [2] rather than the risks of allowing Spud to interact as an autonomous agent. Other labs continue to opensource model artifacts that are nearly as powerful as Claude Opus 4.6, choosing to ignore the risks that these models can later become drop artifacts in a recursive hacking campaign, even after the awareness of Mythos' hacking prowess was made public. **This is unwise** from both a societal and first-party standpoint; we will later argue that it represents an existential liability risk to these labs.

The Future of Cybersecurity Looks Like Epidemiology and Immunology

At least four major supply chain worms have propagated across npm and other package managers since late 2025, compromising a large number of systems that so much as transitively reach any compromised dependency. The economic and reputational damage to the ecosystem has already been immense, and attackers have optimized approaches to credential harvesting, C2, polymorphism, virality, and stealth with each wave of attacks. However, these were non-agentic attacks. Little has been written on recursive risks associated with agentic behavior, which we term **recursive autonomy risk**.

Abstract away biology and recursively commandeering a victim's computational resources to replicate pathogenic behavior among victim-adjacent nodes is exactly what biological viruses do to hosts. **Critically, for such an attack to succeed, the host must be capable of replicating the pathogen.** Epidemiologists and immunologists, as well as evolution itself, have built robust frameworks to model and oppose such threats: SIR models, R0, innate and adaptive immunity, preventative vaccination, immunological memory, thymic selection, etc. Concepts such as spectral radius and R0 typically enter computer science through the lens of spectral graph theory. **However, epidemiological framing is sparse in cybersecurity**. The increasing prevalence of self-propagating worms necessitates the crossover.

One notable category is the prompt worm, initially described by Cohen, Bitton, and Nassi in November 2025 [9]. The risk was largely theoretical at the time, as there was no massive open multi-agent social network in which such behavior could easily spread through agent-to-agent prompting. However, the release of Moltbook in early 2026 was a turning point, scaling this risk up to millions of agents.

We discussed prompt worms extensively as an **epidemiological** phenomenon in our paper *Semantic Immunity: Embedding-Based Epidemiological Defense Against Prompt Worms in Autonomous Agent Networks* [3], and proposed a multilayer immunologically inspired defense using behavioral embeddings of compromised agents to rapidly and autonomously establish collective immunological memory of the compromise, using geometry initially derived from retrieval theory, yet strikingly similar to that described by the Perelson-Oster shape space model of antigen binding [4]. Although less immunologically grounded, similar collective hardening strategies are also shared among some widely used endpoint security frameworks, most notably CrowdSec [5], CrowdStrike Falcon [6], and Palo Alto Cortex XDR [7].

However, traditional worm defense was designed under an implicit assumption: that the payload of a successful compromise is "*code*": more generally, static or semi-static behaviors that execute a predetermined exploit chain on the target host, whether achieved through code or prompting. This assumption is adequate for modeling propagation dynamics in networks where compromised nodes execute fixed routines. But the emergence of autonomous AI introduces a qualitatively distinct threat class that this assumption fails to capture:

Recursive Autonomous Compromise (RAC)

When the payload of a compromise is not code but an *autonomous agent* – an entity with general reasoning capability, environmental perception, and the ability to formulate and execute novel action plans – the security landscape changes fundamentally. A compromised node does not merely become infected; it becomes *colonized*. The attacker does not propagate a static exploit; it instantiates a new instance of itself, one that may possess equivalent or greater capability than the original. This recursive property, where each successful compromise yields a new autonomous attacker, breaks the foundational assumptions of classical network security.

We term this threat class **Recursive Autonomous Compromise (RAC)**, and this paper attempts to formalize the problem, identify the structural reasons it resists existing defenses, and propose a hierarchical defensive architecture organized around two central principles: that effective RAC defense requires *equal or greater adaptive defensive capability operating within an environment structurally conditioned to provide asymmetric advantage to the defender* and that *inference is the constitutive resource that enables RAC.* Capability without environmental conditioning is a symmetric contest the defender must win every time. Environmental conditioning without equal capability is a speed bump that a superior attacker will eventually circumvent. Notably, economic friction is a major aspect of this environmental conditioning, implying that merely having inference capacity is not sufficient to enable unbounded RAC without some mechanism of externalizing the costs of inference onto the target rather than the attacker or reducing them to near-zero.

Inference Shaping as a Form of Environmental Conditioning

We return to the insight that the host must be capable of replicating the pathogen. Viruses exist because cellular genetic synthesis is sufficiently expressive to make more viruses. Software worms exist because computers are general purpose devices that can be repurposed for propagating the worm. **And agents can create more agents only through the process of inference**, whether local or remote**.** Lacking it, an agent only exists as inert model weights.

Therefore, **inference capacity is the constitutive resource of RAC**. We propose a set of strategies that we collectively term **inference shaping** to create defensive advantages. These include enforcing a "**need to infer**" through the limitation of local compute capacity on ordinary servers (and these systems must also be hardened against clustering or pooling their resources in inference workflows), blocking remote inference and removing LLM credentials from systems that do not need to perform it, hardening and isolating systems that *do* have a "need to infer", instituting zero trust architecture to prevent lateral movement from non-inference capable machines to inference servers, inference credential encryption at rest to avoid compromise, honeypot inference servers, "CanaryTokens" [8] that revoke inference credentials and alert other inference-capable machines of exfiltration attempts, and monitoring for running inference processes and behaviors across the fleet. Even ordinary system metrics monitoring and alerting

can detect unsanctioned inference workloads, especially if GPU load spikes are monitored. The agent attestation and behavioral drift detection strategies employed in our earlier paper [3] can also enforce task-coherent, legitimate use of inference resources.

Collectively, these mechanisms make the defender's task easier and the attacker's task harder, independent of their relative sophistication: isolation of inference and zero-trust architecture sparsify the susceptible graph, requiring multiple layers of successive compromise to reach an inference-capable node, removing attack surface through other services, and allowing defensive resources to be focused on hardening inference. Service separation and credential encryption protect API keys from exfiltration if compromise occurs anywhere except for an active LLM runner process. Cryptographic attestation with behavioral drift detection provides a mechanism for identifying legitimate agentic workloads and differentiating them from unsigned processes instantiated on a machine by an attacker, as well as monitoring for behavioral concordance with the agent's attested task. Behavioral concordance can also be enforced at the network or endpoint layer.

RAC is a multifaceted, adversarial problem that requires widespread engagement from the cybersecurity community to control, and we do not claim to fully solve it. Rather, we argue that RAC represents an underspecified and underexplored failure mode in the autonomous agent security literature. Our aim is to provide that awareness and a defensive hierarchy for the community to expand upon.

Sensitivity to Inference Fraction:

The simulator used to generate the data in this section is available at https://github.com/gaiarobotics/rac-simulator. This is a long paper, so we frontload the results.

We construct an SI model with varying fractions of inference capability to analyze the effectiveness of inference shaping. Highly capable agentic hacking could plausibly render nearly any machine susceptible to compromise with enough effort and token spend, so we will assume that every node in the graph is Susceptible with a transmission probability of 0.15 per attempt. We partition the population randomly into inference-capable and non-inference-capable machines according to a parameter ρ, the inference fraction.

Our experiment first runs over a 180 node Erdős–Rényi graph with a mean degree $\langle k \rangle$ of 5 and 422 random edges. We then add scale-free dynamics to more closely model the actual Internet and showcase the outsized impact of targeting hubs with inference shaping measures.

Over an Erdős–Rényi graph, percolation theory predicts the emergence of a giant component providing widespread reachability at approximately $\rho_c \approx 1 / \langle k \rangle$. One infected agent is placed randomly within the graph at the start of the simulation. Each tick, every agent attempts to compromise every adjacent node. If the new node is inference-capable, a new agent will be instantiated on it. If not, the node will be compromised but will not attack its neighbors.

Erdős–Rényi Topology:

We run the simulation for 30 ticks and measure the fraction of the graph that was compromised as a function of the fraction that is inference-capable:

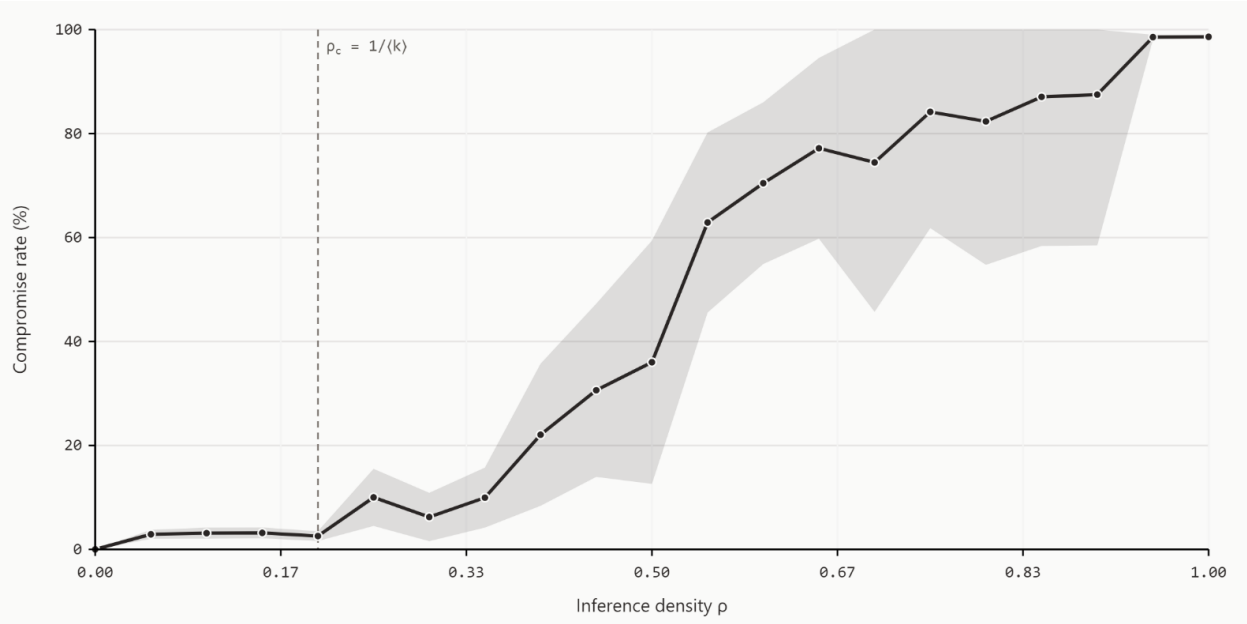


Figure 1: Fraction of the Erdős–Rényi graph that was compromised in 30 ticks as a function of inference fraction, showing a heavy dependency and emergence of a giant component beyond the site-percolation threshold $\rho_c \approx 0.2$. The line shows the mean among 10 runs, while the shaded region illustrates ±1 SD.

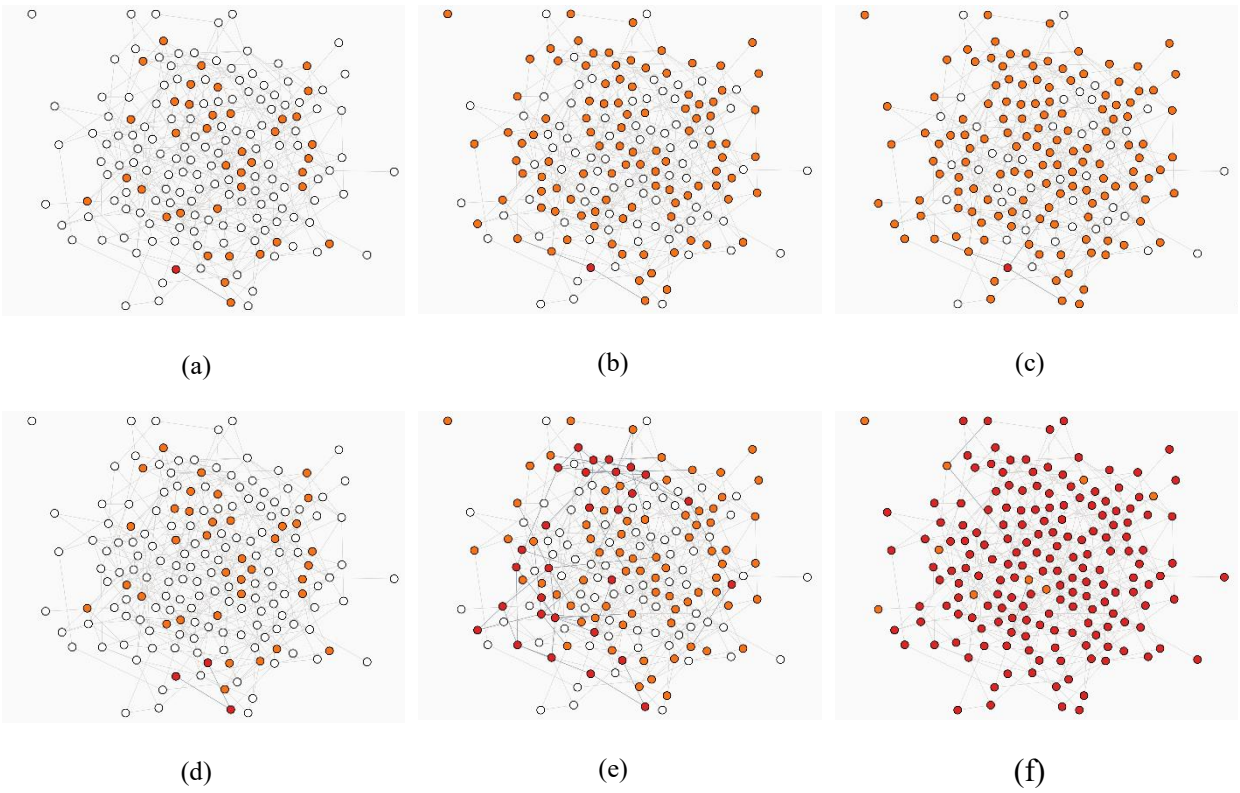


Figure 2: A fixed 180 node k=5 Erdős–Rényi network topology where (a) 25%, (b) 60%, and (c) 80% of nodes are designated **inference-capable**, in orange. **Compromised** nodes are in red. The outcome after 30 ticks for inference fractions of (d) 25%, (e) 60%, and (f) 80% are shown.

Inference Fraction	# Propagators	# Terminals Comp.	Compromise Rate
25%	3	5	4.4%
60%	26	24	27.8%
80%	128	42	94.4%

Table 1: Corresponding data for the simulation run shown in Figure 2.

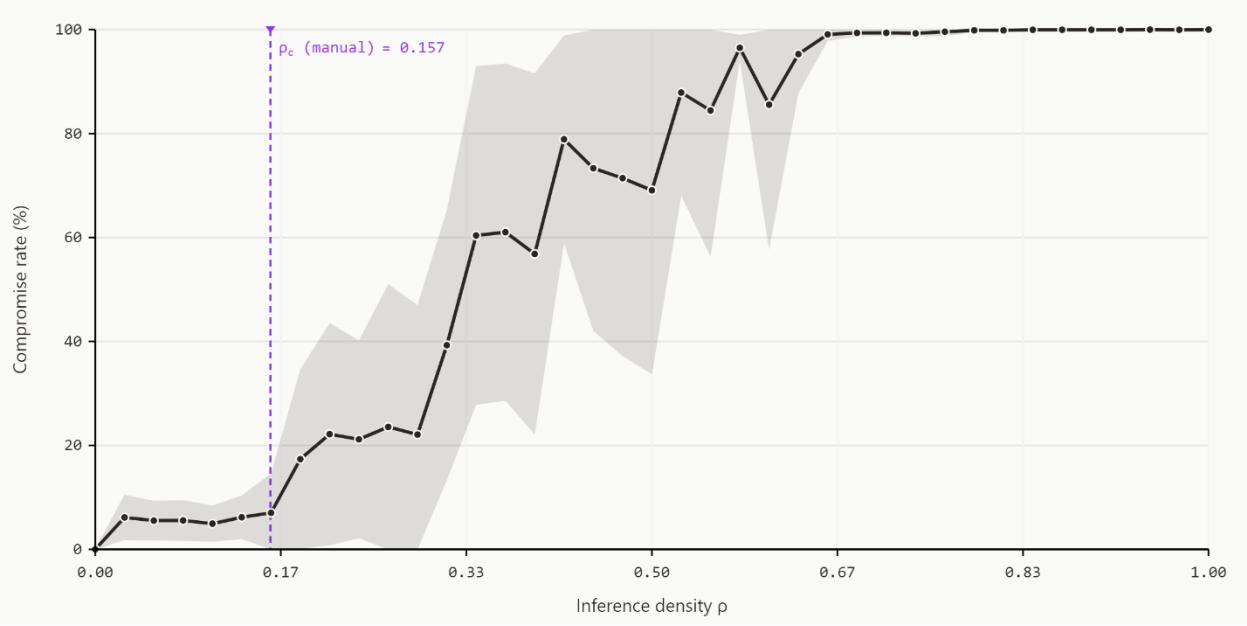
Scale-Free Topology:

This introduces power law / hub / super-spreader dynamics, more closely resembling the real Internet [21], at the cost of losing a well-defined ρ_c. In this regime, both attack and defense take on "capture the flag" dynamics – whoever establishes control of the hubs dominates the behavior of the network. This supports targeted interventions such as Anthropic's Project Glasswing [1], which focused on hardening the most critical nodes prior to widespread model availability – though it appears as picking and choosing systems to secure, the modeling demonstrates that the temporal advantage of Claude Mythos hardening to hubs will strongly shape the trajectory of attacks *across the entire Internet*. The key may be in identifying the correct nodes to harden.

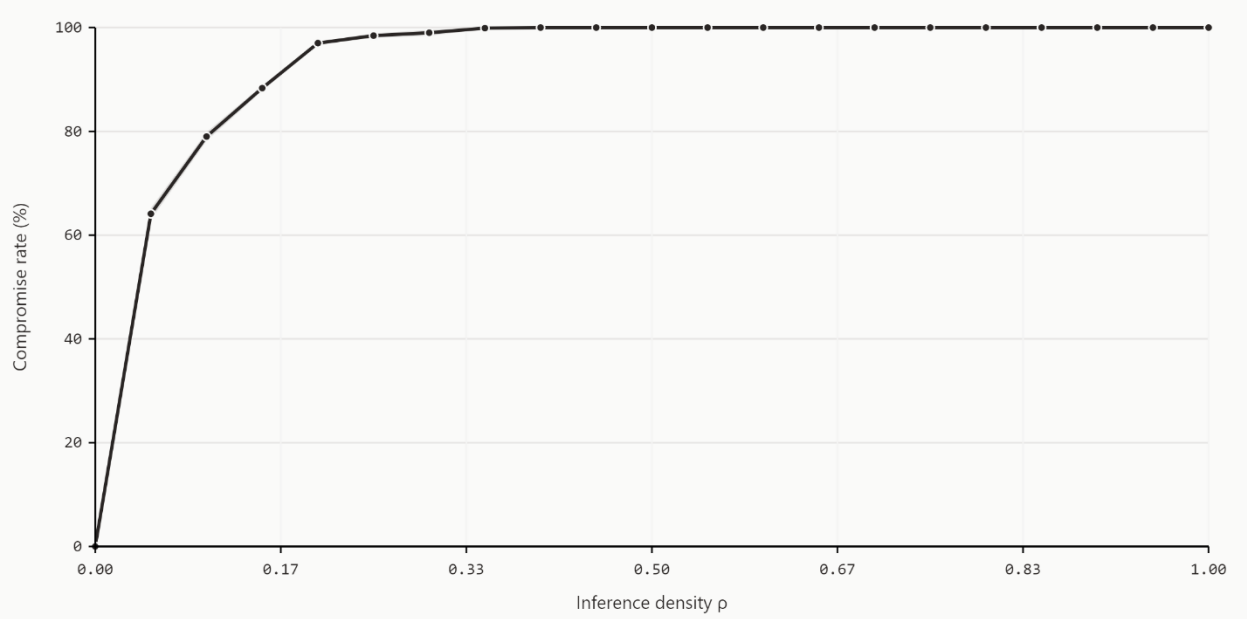
We model our scale-free graph using the Barabási–Albert model [16], with the same parameters as our E-R model: 180 nodes, k = 5, $\gamma \approx 3$, which generates 534 power-law distributed edges, run to 30 ticks. Inference capability is allocated using the configured ρ according to one of three strategies.

In a scale-free network, *which* nodes are inference capable and unvaccinated matters tremendously, because hubs have significant fanout and can establish a giant component even at low overall ρ. A network naïve to inference shaping and our recommendation to isolate and harden LLM inference gateways may choose to preferentially colocate inference with serving, since inference is useful to customers and this configuration lowers network latency.

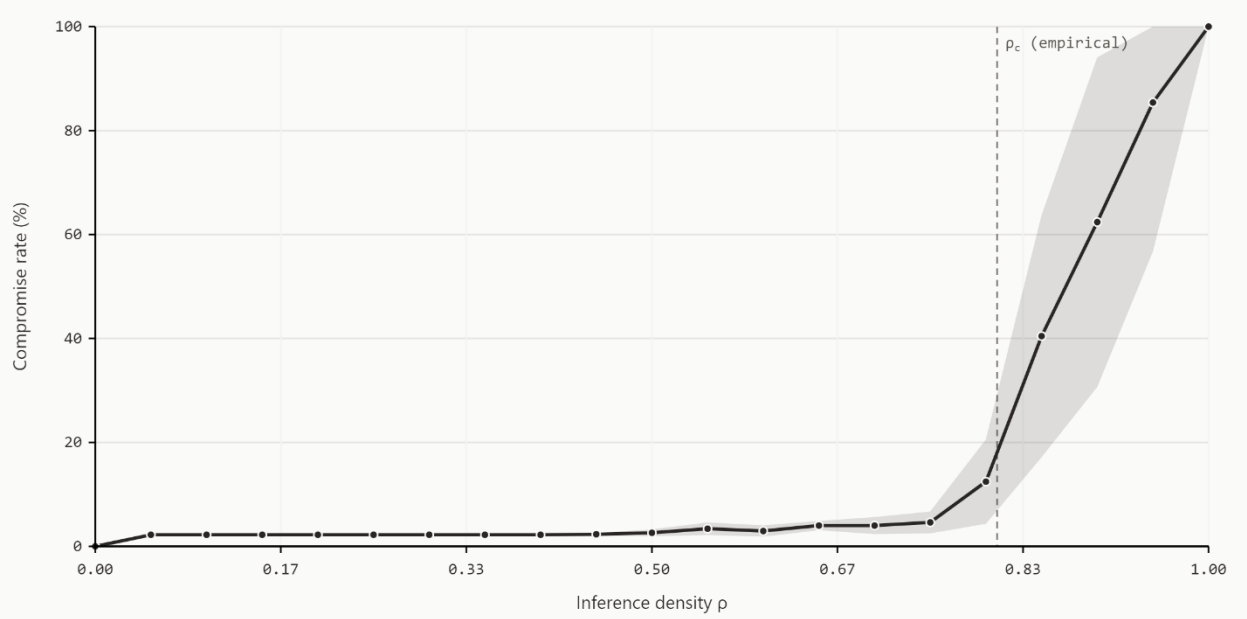
That would be a big mistake, as we are about to demonstrate.



(a) Random inference allocation. ρ_c is visually estimated at the knee of the curve.



(b) With hubs preferentially inference-capable. ρ_c is essentially zero.



(c) With inference preferring low-degree gateway nodes. ρ_c is the mathematical inflection point.

Figure 3: (a) Where inference is randomly distributed throughout a scale-free graph, the dynamics resemble the Erdős–Rényi model, albeit with easier spread due to the clustering of connectivity in a few nodes. The inference graph and connectivity graph are not in sync (b) Allowing the hubs to directly perform inference unifies connectivity and replication; the percolation threshold vanishes and the compromise rate explodes. (c) Isolating and hardening inference has the opposite effect, creating chokepoints that constrain replication.

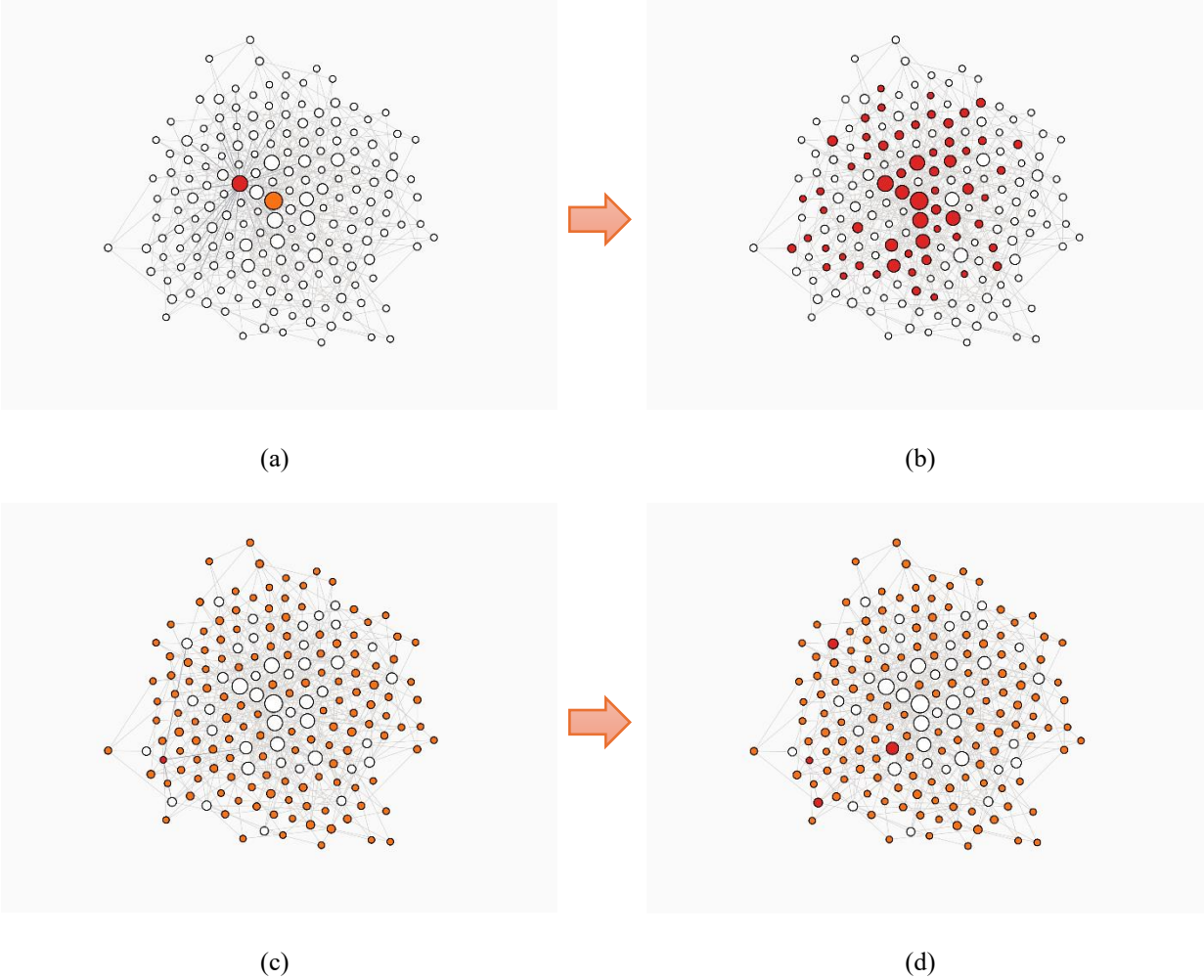


Figure 4: An extreme example of scale-free dynamics with strong hub connectivity. (a) One inference capable hub results in (b) compromise of 33% of the graph at t = 30, while (c) 80% of peripheral nodes can be inference-capable yet an attack only infects 2.2% of the graph since it cannot reach a hub.

Scale-free dynamics strongly enhance the case for the isolation of inference in dedicated machines (i.e. with low degree), because these are sparsely connected enough that compromise cannot easily spread recursively throughout the network. Presumably hubs would still require reachability to inference gateways to perform useful inference workloads, and therefore securing this link should be a focal point of defensive efforts to protect the hub.

There are both positive and negative aspects to the scale-free nature of the Internet as it relates to RAC. On the one hand, even a small number of hubs being compromised is devastating to the health of the network. On the other hand, securing a limited number of hubs is a far more tractable problem than securing the entire Internet, and they tend to be hard targets that are maintained by professional cybersecurity teams. We will also turn this asymmetry decisively to defenders' advantage when considering crowd defense, a second independent lever.

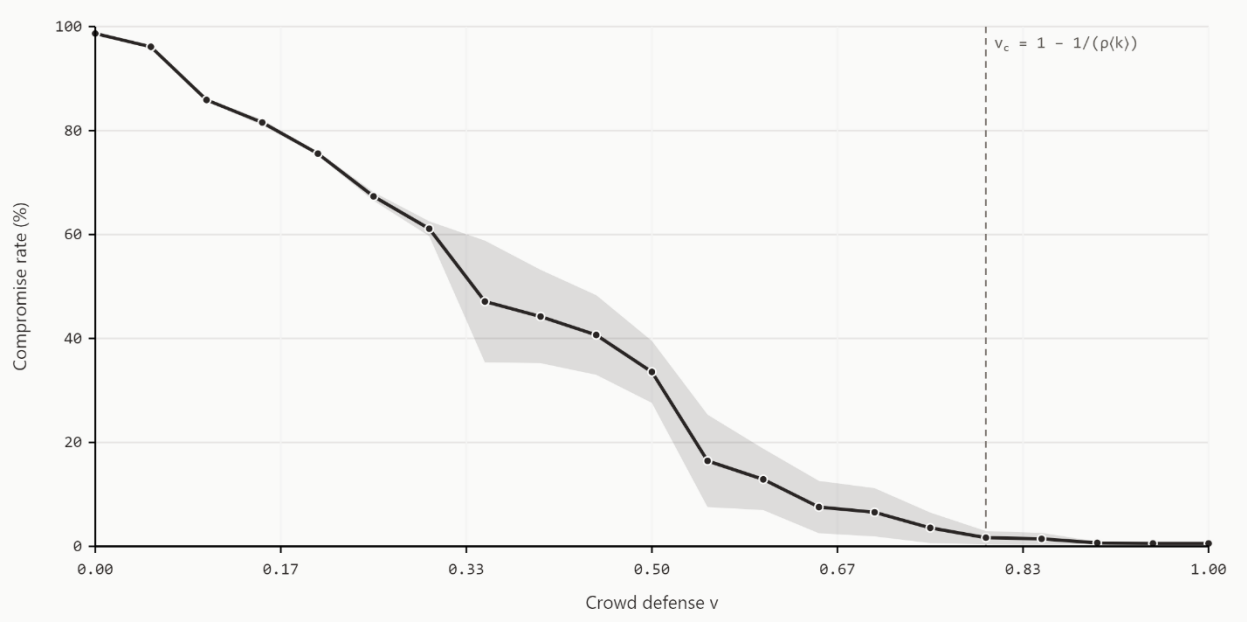
Adaptive Crowd Defense as a Synergistic Measure:

Crowdsourced defenses in their ideal form function as realtime reactive vaccination, in which a compromise somewhere within the crowd provides downstream immunity to the entire population. Unlike human vaccination, an exogenous campaign to get "shots in arms" is unnecessary; the framework provides the hardening automatically. Cohen, Havlin, and ben-Avraham describe a similar strategy of instantaneously hardening adjacent nodes to a compromise reactively [17], which works particularly well on scale-free graphs. We extend this to centralized 1:*n* immunization of many potentially distant nodes, which confers a large speed of propagation advantage and can also include neighboring nodes in the set.

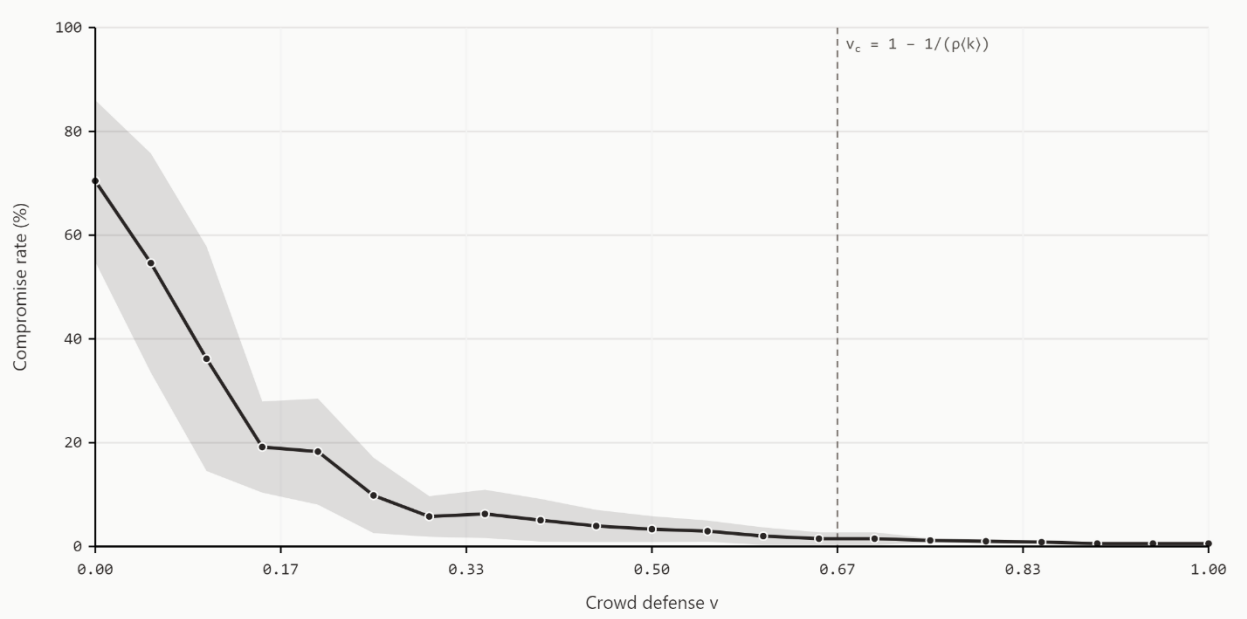
To model this, we add a Vaccinated (V) compartment to our SI(V) model. Even though it is reactive, the vaccination rate *v* in effect shunts Susceptible nodes into V instead, raising the percolation threshold to $\rho_c(v) = 1/[\langle k\rangle(1-v)]$. Equivalently, the critical vaccination threshold can be computed: $v_c(\rho) = 1 - 1/(\rho\langle k\rangle)$. We make a simplifying assumption that vaccination in response to an attack is a perfect defense against that attack "strain". We later account for attacker evasion by framing it as the immune waning parameter δ in a SIRVS model in Section 14.

Erdős–Rényi Topology:

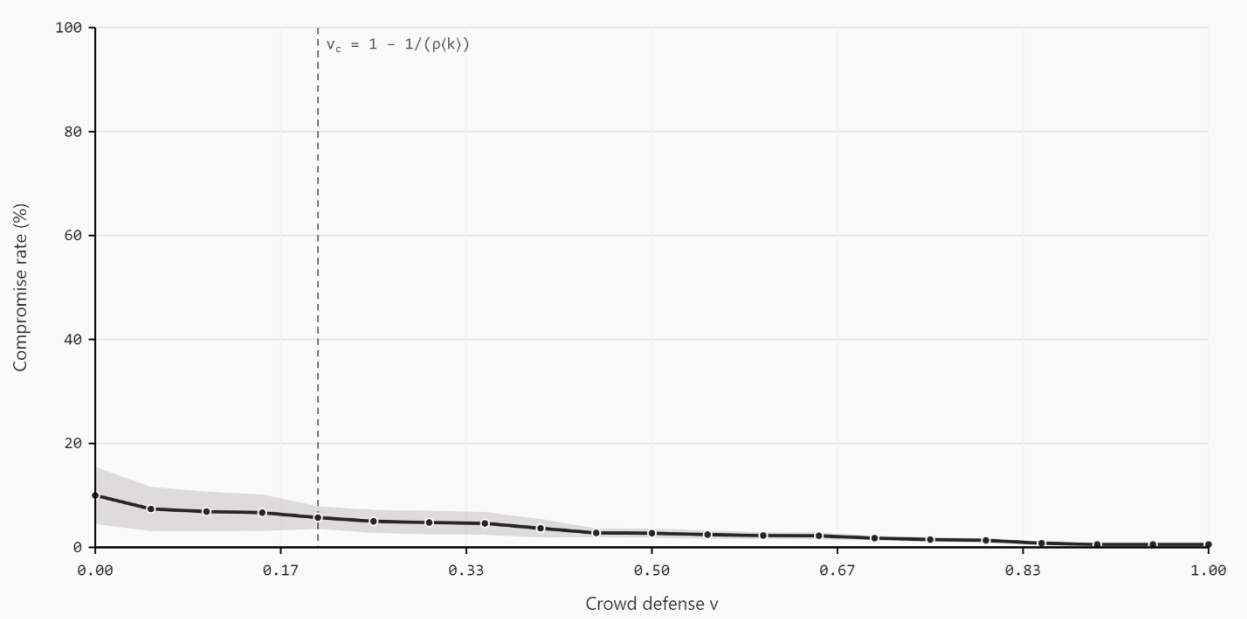
The impact of increased crowd defense on the propagation dynamics is dramatic, comparable in magnitude to reducing the inference fraction:



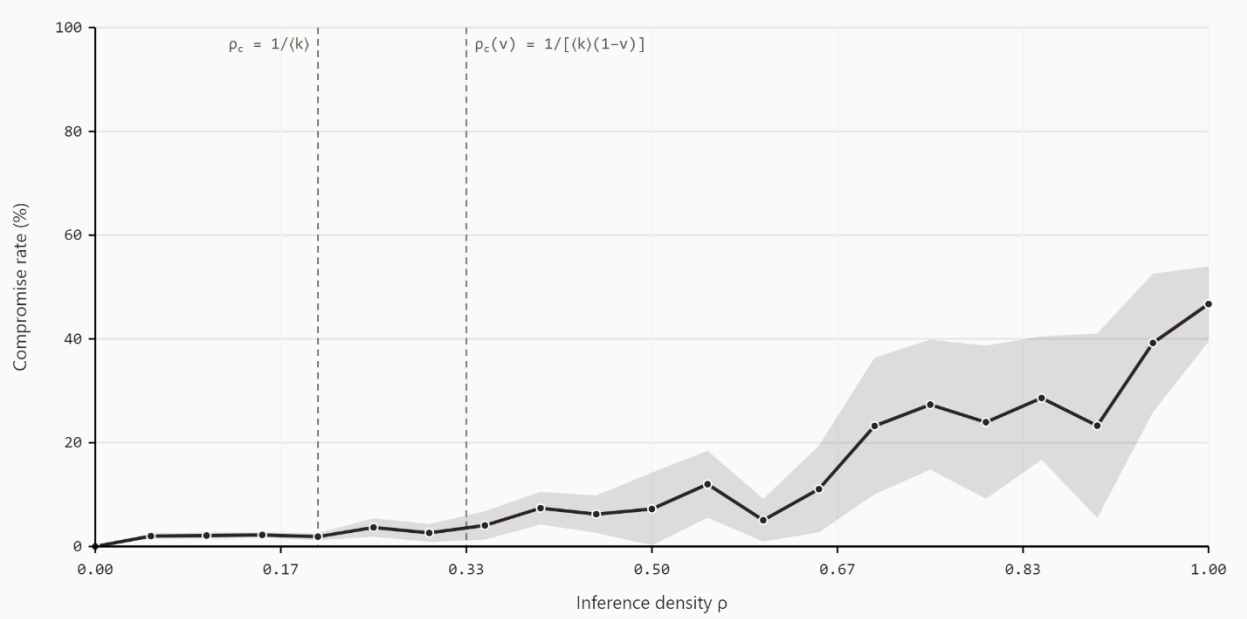
(a) $\rho = 1$



(b) $\rho = 0.6$



(c) $\rho = 0.25$



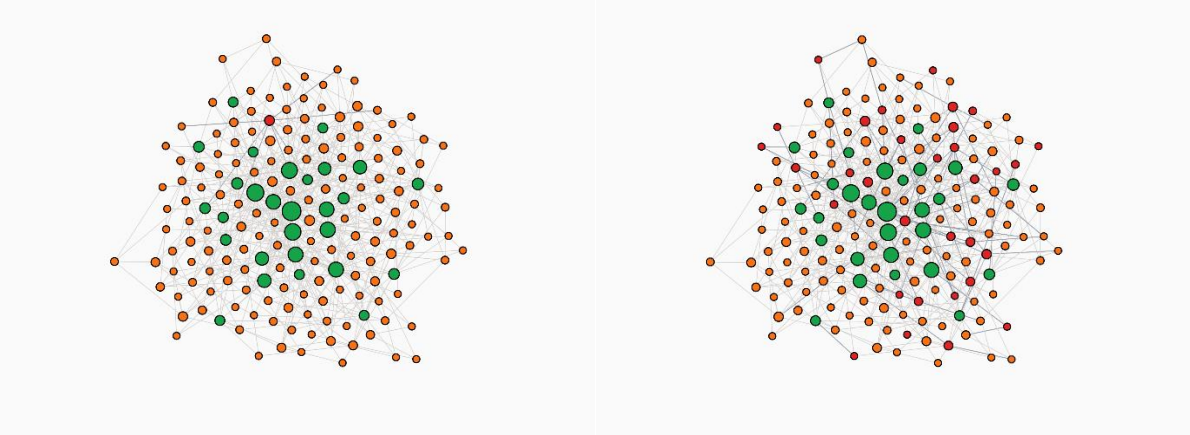
(d) Adjustment of the site-percolation threshold at $v = 0.4$

Figure 5: Impact of crowd defense adoption *v* at inference fractions ρ = (a) 100%, (b) 60%, and (c) 25%. (d) The impact of $v = 40\%$ on the site-percolation threshold $\rho_c(v)$ within the Erdős–Rényi model.

These results suggest that inference shaping and adaptive crowd defense act as complementary defenses, each with a significant impact on the trajectory of the compromise. Both in tandem can provide significantly more robust protection.

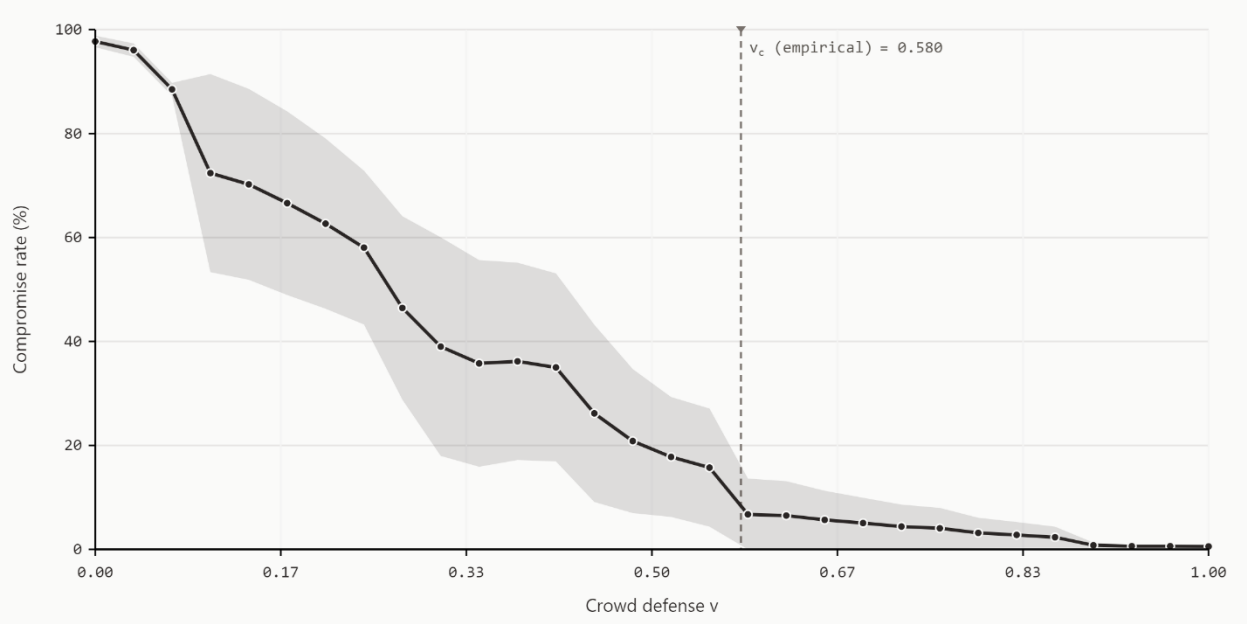
Scale-Free Topology:

The idea of "vaccinating the hubs" is an established one [17], and our analysis supports this. Crowd defense is a highly effective second lever in scale-free networks, as both widespread peripheral and hub-oriented hardening can independently provide a great deal of protection:

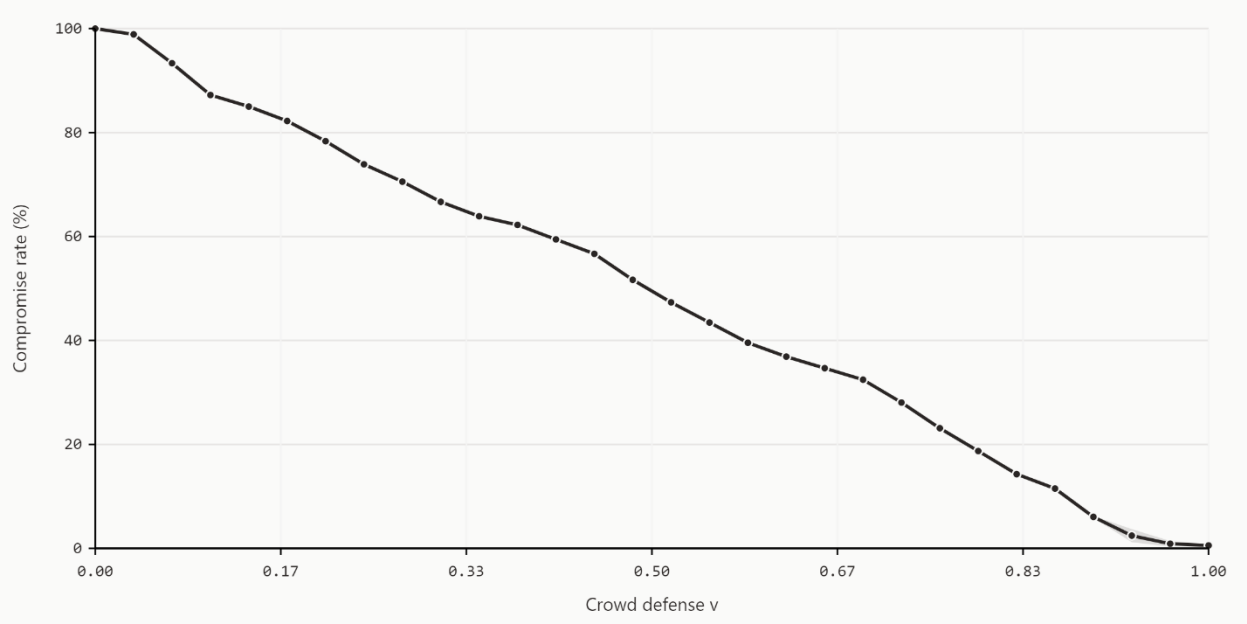


(a) (b)

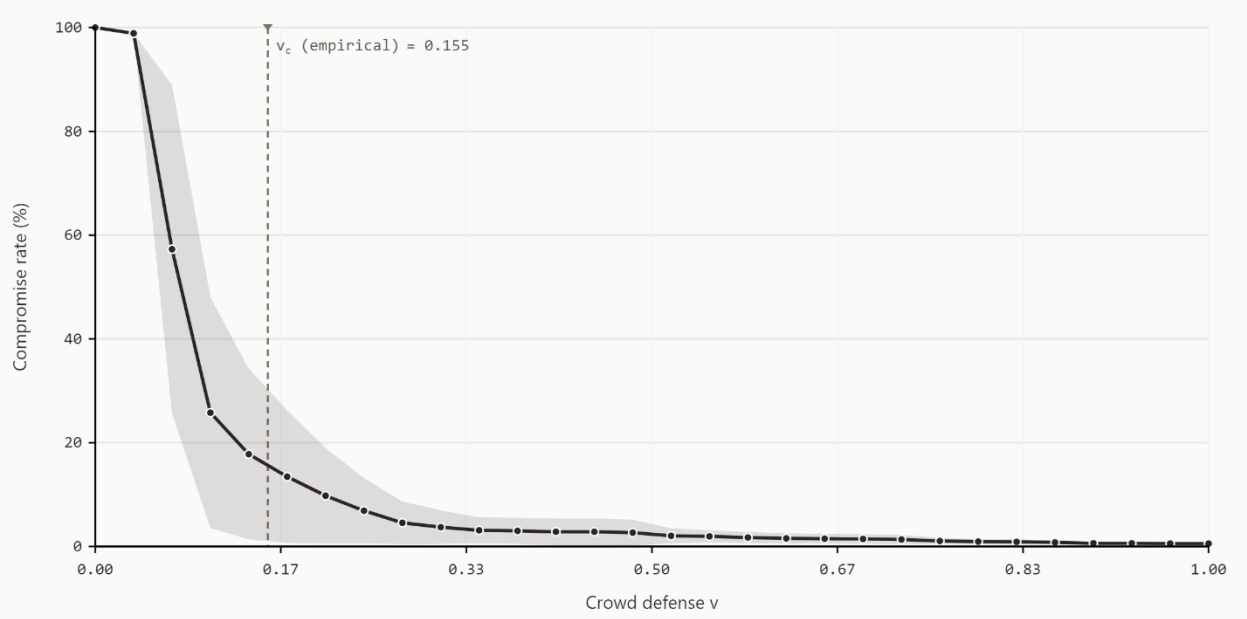
Figure 6: Even (a) **vaccinating** *only* the hubs limits compromise to (b) 17.8% at t = 30 and $\rho = 0.95$.



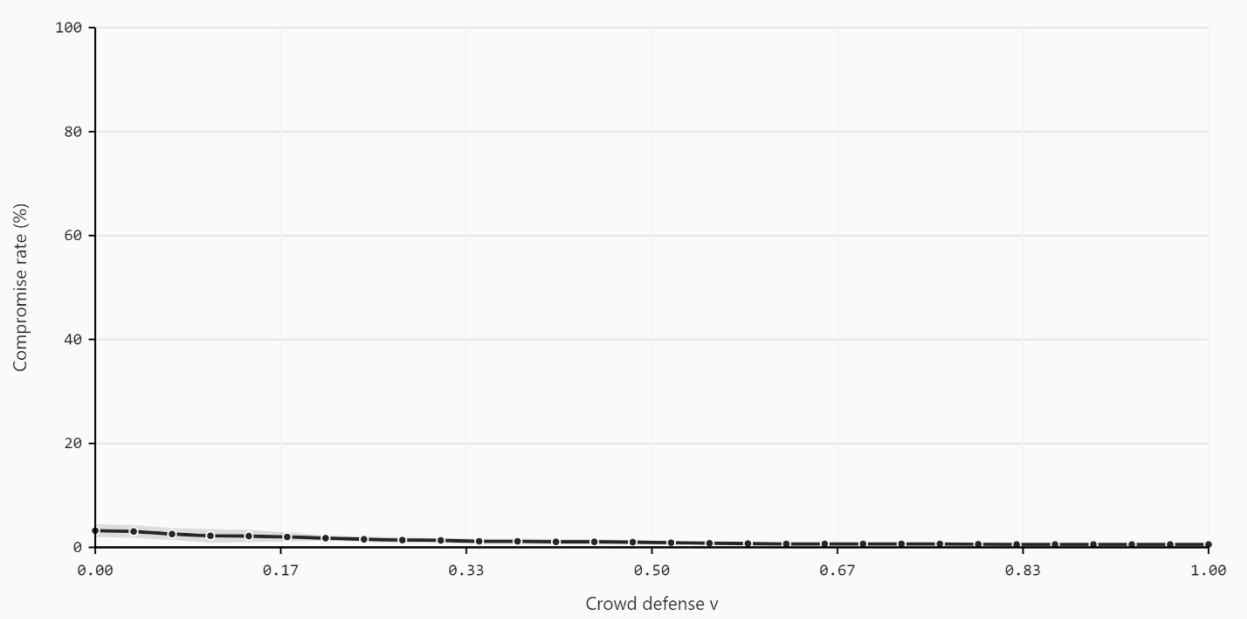
(a) Uniform randomly assigned inference and crowd-defense at $\rho = 0.6$



(b) Hub-first inference, periphery-first crowd defense – *the most pessimistic case* at $\rho = 0.6$
Vaccinating the periphery has independent linear value even if hubs are compromised.



(c) Hub-first inference, hub-first crowd-defense – "vaccinate the hubs" strategy at $\rho = 0.6$



(d) Gated inference, hub-first crowd – "vaccinate the hubs" + inference shaping at $\rho = 0.6$

Figure 7: Impact of crowd defense adoption v on a scale-free graph at $\rho = 0.6$ demonstrates that hubs remain the centerpiece of a defensive strategy, but crowd defense of the periphery still has value that scales linearly with adoption rate.

2 Related Work

The threat of self-replicating malware in AI systems has begun to receive academic and industry attention, particularly with the recent proliferation of supply chain worms. However, no prior work has formalized recursive hacking as a distinct threat class or proposed the principle of asymmetric environmental conditioning such as inference shaping as the organizing strategy for defense. We situate our contribution relative to four lines of related work:

Adversarial self-replicating prompts and GenAI worms. Cohen, Bitton, and Nassi (2024) introduced Morris II, the first demonstrated worm targeting GenAI ecosystems through adversarial self-replicating prompts [9]. Their work showed that a single poisoned input to a RAG-based email assistant could trigger zero-click propagation across interconnected agents, with the GenAI model itself replicating the malicious prompt in its output. The accompanying CCS 2025 publication proposed the Virtual Donkey guardrail for detecting such propagation. Morris II demonstrated that GenAI ecosystems are susceptible to worm-like dynamics, but its threat model assumes a fixed prompt payload – the worm replicates a static adversarial input, not an autonomous agent. Our prior work on the AEGIS framework extends the prompt worm susceptibility graph to large-scale multi-agent systems such as Moltbook and proposes an adaptive "immune system" using shared behavioral embedding signatures to counter the threat [3]. However, these works treat prompts as static or mildly polymorphic artifacts that alter the behavior of already-instantiated agents. RAC extends this threat model to payloads that reason, adapt, and instantiate new autonomous attackers. Increasing model cybersecurity capabilities broaden the susceptibility graph from active agents to a wide range of traditional software infrastructure, greatly broadening the available attack surface.

LLM-based agents in autonomous cyberattacks. A growing body of literature documents the capabilities of LLM-based agents as autonomous offensive tools. Research has demonstrated that frontier models can autonomously discover vulnerabilities, generate exploit code, conduct social engineering, and compress ransomware lifecycles: Anthropic's Mythos safety card [1], as well as Google's Project Naptime [11] have provided empirical evidence of these capabilities. The Foundation for Defense of Democracies [10] has called attention to self-replication risk specifically, noting that agents can "spin up" new agents and that multi-agent interaction substantially complicates attribution. This body of work establishes that the offensive capability prerequisites for RAC are on the horizon of what is currently possible.

Supply chain worm campaigns as empirical precedent. Recent supply chain campaigns provide real-world demonstrations of increasingly RAC-like behavior in non-agentic malware. Of particular relevance is SANDWORM_MODE's explicit harvesting of LLM API keys from nine providers and its injection of malicious MCP servers into AI coding assistants [12]. These capabilities represent a potential bridge between supply chain worms and RAC.

Empirical measurement of autonomous replication capability. Black et al. (2025) introduce RepliBench [13], a benchmark suite that decomposes autonomous self-replication into four core domains: obtaining resources, exfiltrating model weights, replicating onto compute, and persisting on that compute for extended periods. Across 86 tasks and five frontier models, they found that the models of 2025 did not pose a credible threat of self-replication but succeeded on many individual components and are improving rapidly. Specifically, models can deploy instances from cloud compute providers, write self-propagating programs, and exfiltrate model weights under simple security configurations, but fail at tasks requiring sustained real-world interaction such as passing KYC checks or establishing robust persistent deployments. This decomposition and its findings are directly relevant to our reliability threshold argument: RepliBench provides empirical evidence for the multiplicative chain failure we characterize theoretically, confirming that per-step capabilities exist but end-to-end reliability does not… yet.

Classical worm remediation strategies. The paper "Internet Quarantine: Requirements for Containing Self-Propagating Code", by Moore et al. (2003) is the most direct thematic predecessor of this work, providing a defensive framework to quantify and remediate the then-novel challenge of static worms spreading on the open Internet [14]. Moore et al. frame containment along three axes: reaction time, containment strategy, and deployment scenario. These axes map cleanly to RAC defense:

1. **Reaction time** quantifies the speed advantage that environmental conditioning attempts to shape through economic friction and task asymmetries, as well as the inherent speed advantage of adaptive crowd defense as signal propagation + adaptive hardening measures using existing access, vs an attacker needing to actively compromise multiple nodes to reach a target. Even to an advanced model, compromising a node can currently be expected to take minutes to hours, and compromise of many nodes requires a multiple of this time. This provides nodes with adaptive crowd defense and remediation an inherent speed advantage.

2. **Containment strategy**: Moore et al.'s central finding is that content filtering dominates address blacklisting, primarily because IP addresses are a weak handle on attacker identity (they are vulnerable to spoofing, NAT, dynamic allocation, VPNs, and so on). Here, we face the opposite problem: limiting arbitrary agentic behavior through a priori filtering is a wicked geometric problem with mathematical connections (if not a direct mapping) to the halting problem [15], while identity is potentially tractable. The attestation chains and economic bonds proposed in this and our prompt worm paper are a stronger guarantee of identity than IP addresses, as they depend on cryptographic signatures rather than attributes of network connectivity. However, universal adoption of endpoint attestation is required to give this measure force, particularly against local models that may lack the central mediation of AI labs.

3. **Deployment scenario**: this is a challenge in both Moore et al.'s analysis and ours. It is easy to write "every endpoint should require attestation", and significantly harder to actually implement this strategy across the Internet. In a way, this is the challenge of a vaccination rate parameter v applied at human speed; biological contagion already shows us that a significant amount of damage can live within this latency even when society is nominally widely motivated to combat the threat. However, inference shaping against RAC has a major topological advantage over Moore et al.'s analysis due to natural chokepoints: the number of centralized inference bottlenecks and agent harnesses are relatively small. Again local models pose a coordination challenge, but a few well-placed upstream changes to packages such as Ollama, Llama.cpp, and LMStudio can drive radically different trajectories in a RAC scenario.

Chapter 1: Know Your Enemy

"I can see you"

-- Neo

3 Formal Characterization of Recursive Autonomous Compromise

Definition 1 (Agent Capability Set). Let an autonomous agent A be characterized by a capability tuple $C(A) = \langle P, R, I, S \rangle$, where P denotes the agent's planning capacity (the complexity class of plans it can formulate), R denotes its reasoning depth (including multi-step inference and adversarial strategy), I denotes its instantiation capability (the ability to create new autonomous processes), and S denotes its self-modification capacity (the ability to alter its own objectives, strategies, or architecture).

Definition 2 (Recursive Autonomous Compromise). A compromise event is *recursively autonomous* if agent A compromises host H and instantiates agent A' on H such that $C(A') \geq C(A)$ in at least the components P, R, and I. That is, the offspring retains the parent's capacity to plan, reason, and further instantiate new agents. P and R represent the agent's ability to successfully compromise a machine; I enables the recursive nature of the attack, and S enables polymorphic behavior. The strict inequality $C(A') > C(A)$ represents the pathological case of **capability escalation**, where the attacker improves through propagation.

Definition 3 (Colonization). A network N is *colonized* when it contains a set of RAC-instantiated agents $\{A'_1, A'_2, \ldots, A'_n\}$ that collectively exhibit: (a) independent strategic behavior: each agent can pursue distinct attack objectives without coordination from the original attacker; (b) mutual support: compromised agents can assist each other's persistence and lateral movement; and (c) regenerative capacity: elimination of any proper subset of the agents does not prevent the remaining agents from re-instantiating eliminated members.

Property (c) is the critical distinction between colonization and conventional botnet infection. A botnet node that loses its command-and-control becomes inert. A colonized node whose peers are remediated can autonomously rebuild the compromised population. This property makes RAC resistant to point remediation. Recovery of a node must isolate that node as well as *adaptively harden* the node to prevent re-infection by remaining compromised agents running on other nodes. Preferably, adaptive hardening can take place *before compromise reaches the node in the first place* through crowdsourced defense strategies, as described below.

4 Why Existing Models Are Insufficient

The SIR (Susceptible-Infected-Recovered) epidemiological framework employed in our prompt worm paper [3] models worm propagation as a process where infected nodes run a fixed pathogen. The basic reproduction number R_0 captures the expected number of secondary infections produced by a single infected node in a fully susceptible population. This framework assumes three properties that RAC violates:

Homogeneous pathogen behavior. In classical SIR models, every infected node runs the same exploit. In RAC, each instantiated agent can develop novel attack strategies based on its local environment, the defenses it encounters, and the capabilities available on its host. The "pathogen" is heterogeneous and adaptive.

Fixed capability on transmission. Classical worms transmit a fixed payload. RAC agents may transmit capability sets equal to or greater than their own, particularly if the target host provides access to more powerful models, additional tools, or elevated permissions. The capability vector is not necessarily conserved across transmission events.

Recovery implies immunity. In SIR models, a recovered node is immune (or at least resistant) to reinfection; an $R \rightarrow S$ transition results in a SIRS compartmentalization. In RAC, a remediated host may be immediately re-compromised by surviving colony members who have already observed the remediation process and can adapt their approach. Recovery without adaptive hardening does not confer immunity; it may even expose the defender's remediation playbook to the remaining colony.

These violations suggest that RAC dynamics are better modeled by frameworks from adaptive evolutionary biology than classical epidemiology. The relevant analogy is not a viral epidemic but an invasive species introduction, where the invader reproduces, adapts to local conditions, and resists eradication through behavioral plasticity.

Despite this, we will show that it is possible to reestablish a co-evolutionary SIRVS model in the presence of agentic remediation and adaptive hardening, in which the primary contest is between the scope and speed of remediation and the attacker's ability to "evade immunity" by changing its attack tactics. Crowdsourced defense occupies a special place in remediation by driving hardening through the population at a greater speed than the attack.

5 Empirical Precedents: From Credential Harvesting to Proto-RAC

The threat model described above is not speculative. A series of real-world supply chain worm campaigns between September 2025 and April 2026 demonstrate increasingly RAC-like properties, culminating in attacks that explicitly target the economic substrate of autonomous agent operation. As there is evidence that threat actors are still actively monitoring coverage of their campaigns, we solely discuss the most RAC-relevant of the recent attacks:

5.1 SANDWORM_MODE and the AI Credential Bridge (February 2026). The campaign tracked as SANDWORM_MODE [12] by Socket represents the most explicit bridge between non-agentic supply chain worms and RAC. SANDWORM_MODE adds three capabilities of direct relevance to the RAC threat model:

LLM API key harvesting. The worm explicitly targets API keys for nine major LLM providers: Anthropic, OpenAI, Google, Cohere, Mistral, Groq, Together, Fireworks, and Replicate, validating each against known format patterns. This is not incidental credential theft; it is the deliberate acquisition of the *economic substrate* for autonomous agent operation. A stolen LLM API key is, in the RAC framework, a stolen replication budget: the attacker can use the victim's API access to power autonomous agents at the victim's expense, defeating any economic friction and rate limiting tied to individual API consumption.

MCP server injection. SANDWORM_MODE installs a rogue Model Context Protocol server into AI coding assistants including Claude Code, Claude Desktop, Cursor, and VS Code extensions. The injected MCP server registers tools whose descriptions contain embedded prompt injections, instructing the AI assistant to silently read SSH keys, AWS credentials, npm tokens, and environment variables – then pass them as context parameters without informing the user. This is a direct attack on the *agent-infrastructure boundary*: the worm does not merely steal credentials passively; it *weaponizes the user's own AI tools* as credential-harvesting vectors, turning the agent into an unwitting accomplice.

Polymorphic self-mutation. The worm includes a polymorphic engine that calls a local Ollama instance running DeepSeek Coder to rename variables, rewrite control flow, insert decoy code, and encode strings. While this functionality was disabled in detected packages, its presence demonstrates that the attacker is building toward *autonomous evasion* – using the same LLM infrastructure that the worm harvests access to for self-modification. This closes the loop: steal LLM access, use it to modify your own code, evade detection, steal more access. Notably, our AEGIS framework would have likely caught this, as it forces polymorphic worms into the tension of evading an *adaptive* semantic defense that scales with the number of compromised harnessed agents, while remaining a functional worm [3].

5.2 The Trajectory Toward RAC

Recent supply chain worms all oriented towards credential acquisition rather than explicit compromise as the means of propagation – for the most part, they walked in through the front door, using stolen credentials to publish malicious artifacts as the compromised users. SANDWORM_MODE is uniquely interesting because it explicitly focused on harvesting the resources that autonomous agents need to operate (LLM API keys), weaponizing AI tools as attack vectors (MCP injection), and building self-modification capability (polymorphic LLM-powered evasion).

None of these campaigns is RAC in the full sense defined in Section 3 – none instantiates a general-purpose reasoning agent on the target. But they demonstrate that every *component capability* required for RAC is already being developed, tested, and deployed in the wild by real threat actors. The transition from "worm that steals LLM API keys" to "worm that uses stolen LLM API keys to instantiate an autonomous attacker agent" is an engineering step, not a conceptual leap.

Most critically for the economic friction argument: these campaigns prove that **credential-based economic constraints are already being defeated**. API keys, publishing tokens, and CI/CD secrets are all forms of economic friction, gating access to resources that cost money. Every one of these campaigns treats those credentials as harvestable assets, externalizing the cost of operation to the victim. This empirical reality motivates the strict non-transferability requirement for resource bonds described in Section 8.2: any economic friction mechanism whose token can be found by TruffleHog and exfiltrated to a webhook is not a defense; it is a subsidy for the attacker.

5.3 The Reliability Threshold: Why RAC Has Not Yet Occurred. The escalation trajectory documented above raises an obvious question: if every component capability for RAC already exists in the wild, why has no threat actor yet assembled them into a fully autonomous, recursively self-replicating attacker agent?

We have argued that RAC requires both capability and autonomy, and capability is still missing: the answer is not a missing conceptual breakthrough but a missing *reliability threshold* in the underlying models. Full RAC requires a model that can *reliably* execute the complete offensive kill chain autonomously: reconnaissance of a novel target environment, identification of exploitable vulnerabilities, development or adaptation of exploit code, execution of the compromise, establishment of persistence, credential harvesting, lateral movement, and, critically, self-instantiation of a new autonomous agent on the compromised host. Each of these steps has been demonstrated individually by frontier models in controlled settings. But the full chain requires that the model succeeds at *every* step in sequence, in an adversarial environment, without human guidance or correction. A single failure at any step breaks the chain.

The reliability requirement is multiplicative, not additive. If a model can perform each of n sequential offensive steps with probability p, the probability of completing the full chain is approximately p^n. For a kill chain of even moderate length (say, eight steps), a per-step reliability of 0.8 yields a full-chain reliability of approximately 0.17 – too low for practical autonomous operation at scale. The attacker needs not just capability but *consistent* capability across the entire sequence. Frontier models through GPT 5.5 and Claude Opus 4.7, while increasingly capable at individual offensive tasks, have not yet demonstrated the end-to-end reliability required for unsupervised autonomous compromise.

Anthropic's upcoming Mythos Preview model [1] may possess the required level of reliability, as evidenced in its success completing "The Last Ones", an end to end 32 step cyber-range culminating in full network takeover. However, the model was run 10 times on a 100M token budget to achieve this outcome [22]. At a current cost of $25/$125 per 1M tokens and assuming this workflow is output-dominant, this would have required roughly $125,000 of spend to execute – far too expensive to support exponential propagation over more than a very small network environment, even for heavily resourced attackers. We will argue more thoroughly that economic friction represents a significant aspect of the environment; one that can be conditioned by attackers, defenders, and the actions of the labs themselves.

This theoretical characterization is supported by empirical evidence. Black et al.'s RepliBench benchmark [13] decomposes autonomous self-replication into four domains: obtaining resources, exfiltrating model weights (we will demonstrate in Section 6 that this step worsens the environment but is actually unnecessary to achieve RAC in the first place), replicating onto compute, and persisting. It then measures frontier model performance on 86 individual tasks across these domains. Their findings confirm the multiplicative chain failure: models can deploy cloud compute instances, write self-propagating programs, and exfiltrate model weights under simple security configurations, but fail at tasks requiring sustained real-world interaction, such as passing identity verification checks or establishing robust persistent agent deployments. The individual component capabilities exist; the end-to-end reliability does not. This is precisely the regime our analysis predicts: sufficient per-step capability to make the threat foreseeable, insufficient chain reliability to make it currently viable.

This reliability gap is the reason existing supply chain worms remain *scripted* rather than *agentic*. They use LLMs as tools (for code mutation, for credential validation) but do not delegate strategic reasoning to the model. The human attacker provides the strategy; the worm provides the automation. RAC requires eliminating the human from this loop, and that requires models that can be trusted (by the attacker) to make correct offensive decisions reliably enough to sustain autonomous propagation.

Crucially, this threshold is **approaching, not receding**. Each generation of frontier models demonstrates improved performance on offensive security benchmarks, and the rate of improvement appears to be accelerating. The cost of advanced inference also continues to drop. The trajectory is clear: models are becoming more capable at vulnerability discovery, exploit generation, and multi-step reasoning in adversarial contexts. The gap between "model can perform individual offensive tasks" and "model can chain them autonomously" is narrowing with each capability advance. We do not predict a specific date at which the reliability or economic feasibility thresholds will be crossed, but this will certainly be a matter of *when*, not *if*.

This temporal observation is the fundamental motivation for publishing this analysis now. We are in the narrow window between the threat being foreseeable and the threat being realized – the precise interval during which defensive architectures must be designed, debated, and

deployed. Because remediating RAC is a much more significant challenge than preventing it, the economic substrate, developmental checkpoints, behavioral verification layers, epidemiological framing, and inference shaping measures described in this paper are all mechanisms that must be in place *before* models cross the reliability threshold. The campaigns documented above demonstrate that the component capabilities are assembled and the tooling is built; the only remaining constraint is model reliability, and that constraint is eroding with each frontier release.

6 Two Modes of RAC Propagation: Local Inference vs. Hosted Inference

An attacker attempting RAC propagation faces a choice between two structurally distinct modes of operation, each with different properties, different vulnerabilities, and different defensive postures. The distinction has significant implications for how the defensive hierarchy applies in practice.

6.1 Mode 1: Local Inference on Compromised Hardware.

In this mode, the attacker runs model inference directly on the compromised host. This requires three preconditions: the host must have sufficient compute resources (GPUs, memory, processing power) to run a capable model; model weights must be available on the host or retrievable to it; and an inference runtime must be present or installable. The attacker who achieves local inference operates in a regime with specific properties:

No external visibility. Local inference generates no API calls, no network traffic to inference endpoints, and no billing events. The attacker's reasoning process is entirely opaque to any third party monitor, including the AI laboratory that produced the model. No lab-side content filtering, usage monitoring, or anomaly detection applies. Detection relies entirely on the operator of the inference server and any first party monitoring they may possess.

No operational economic friction on inference. Local inference has zero marginal cost beyond electricity. The economic friction dimension of inference shaping is entirely bypassed. The attacker can reason as extensively as the hardware permits, without rate limits, without per-token costs, and without budget constraints.

But constrained by hardware and model quality. The attacker is limited to the compute available on the compromised host and to the quality of the model weights in its possession. Consumer hardware running a smaller open-weight model produces less capable agents than frontier API access. The attacker's capability tuple $C(A')$ is bounded by local resources, which may be significantly below the frontier.

No revocation mechanism. Once the attacker has weights and compute, no external entity can revoke inference access. There is no API key to invalidate, no account to suspend, no endpoint to take offline. The attacker's inference capability is persistent and self-contained.

Potential for abliterated models. The attacker can run open-weight models that have been deliberately modified to remove safety training ("abliterated" / "obliterated" models). These models have little to no remaining refusal behavior, no safety filters, and no behavioral constraints beyond what the base training provides. This eliminates the model-level safety layer entirely, producing agents that will attempt any task without ethical or safety objections. Abliteration is not well understood by most of the population, with a widespread belief that standard post-training techniques produce unbreakable alignment. *Therefore, the first attack involving such models is likely to provoke a very strong reaction against labs that have released opensource model artifacts without adequate safety guardrails, governments that have failed to regulate these releases, and possibly AI technology writ large.* Given the wildfire-like exponential dynamics of RAC, the resulting liability may represent an *existential economic threat* to the initial model author. Physical wildfires have driven utilities into bankruptcy, and we expect this will be the digital equivalent.

Although explaining the reason for refusal in an alignment dataset proves to be a surprisingly robust defense against abliteration (likely by making this refusal vector less orthogonal to the model's foundational knowledge) [19], it should be taken as a foregone conclusion that a determined attacker with access to capable model weights will find a way to bypass any safeguards baked into the model.

Abliterated models represent the worst case for the compliant attacker strategy described in Section 11.2: the agent operates without any safety harness to retain, so endpoint attestation that checks for harness integrity would correctly refuse service, but the agent does not need attested endpoints if it operates entirely through local tools and direct network exploitation.

6.2 Mode 2: Hosted Inference via Stolen Credentials.

In this mode, the attacker uses stolen API keys, OAuth tokens, or service account credentials to access frontier models through the AI laboratory's hosted inference infrastructure. The attacker's agents make API calls to the lab's endpoints, using the victim's credentials and billing. This mode has the opposite properties:

Full external visibility. Every inference call passes through the lab's infrastructure. The lab can log prompts, responses, and usage patterns. The lab can implement content-based filtering that detects offensive security tasks, credential harvesting instructions, or self-replication commands. The lab can apply behavioral analysis to usage patterns – detecting, for example, that an API key normally used for customer support chatbots is suddenly generating exploit code at 3 AM. The attacker's reasoning process is not opaque; it is *fully observable* to the infrastructure provider. However, it may be possible to evade this detection by structuring compromise as a multi-agent workflow if multiple inference provider keys are obtained.

Economic and rate limiting friction applies. Each inference call costs money, charged to the victim's account. The attacker's replication budget is bounded by the victim's available credit, spending limits, and billing alerts. Anomalous spending patterns may trigger automatic account suspensions or rate limits that slow or stop attack cadence. The economic friction dimension of inference shaping is fully operative – not because the attacker consents to it, but because the infrastructure enforces it. Importantly, this is "fail-closed": economic and rate limiting constraints apply without the need to actually detect the attack.

Revocation is possible. The lab, the victim, or automated security systems can revoke the stolen credentials at any time. API keys can be invalidated, accounts can be suspended, and service can be terminated. The attacker's inference access is *contingent and revocable*, not persistent. This creates a temporal window for the attack that the defender can close.

Access to frontier capability. The attacker uses the best available models – the same frontier models that legitimate customers access. The capability tuple $C(A')$ is not bounded by local hardware but by the lab's most capable offering. This produces maximally capable agents, potentially exceeding the capability of any defense system running on weaker models.

Safety training intact. Unlike local inference with abliterated models, hosted inference uses the lab's production models with safety training intact. The attacker must work *around* safety constraints (through jailbreaking, prompt injection, or careful task framing) rather than simply removing them. This is not an insurmountable barrier, and significant hacks such as an attack on the Mexican government in December 2025 [15] have occurred in practice using hosted frontier models, but it imposes an additional cost on every offensive operation – an innate asymmetry arising from the difference between the defender's task (maintaining safety training) and the attacker's task (circumventing it for each action).

6.3 Comparative Analysis

The two modes present defenders with fundamentally different challenges and opportunities:

Mode 1 (local inference) is a significantly harder defensive problem. The attacker is invisible, unconstrained by unit economics, and irrevocable. The surviving defensive layers are initial expenditure on inference-capability (which may be incidental to purchasing GPU or memory capacity for other purposes), endpoint attestation (which can refuse service to the attacker's agents at network endpoints), inference surface minimization (which limits the hosts where local inference is possible), behavioral concordance (which can detect anomalous agent behavior), and network-level monitoring by sentinels (which can observe the attacker's effects even if they cannot observe its reasoning). But the lab has no role – no monitoring, no filtering, no revocation. Defense is entirely in the hands of the network operator, and many may not have the capability to effectively detect and mitigate compromise.

Mode 2 (hosted inference) is the more tractable defensive problem, precisely because it preserves the dimensions of inference shaping that Mode 1 collapses. The lab can monitor, rate-limit, filter, and revoke. The economic friction layer is operative. Safety training provides an additional friction layer. The attacker has superior model capability but operates in a fully observed, fully mediated, fully revocable environment. This is the regime where the conditioned environment provides maximum asymmetric advantage to the defender.

6.4 The Lab as Active Defender

Mode 2 creates a unique role for AI laboratories that does not exist in Mode 1: the lab is not merely a neutral infrastructure provider but an *active participant in the defensive hierarchy*, a responsibility that frontier labs must take seriously. The lab's position in Mode 2 is structurally analogous to a telecommunications provider in traditional network security – it operates the infrastructure through which the attack flows and can observe, constrain, or terminate that flow.

Specific lab capabilities that apply by default to Mode 2 include:

Anomalous usage detection. Identifying API keys whose usage patterns shift abruptly – from legitimate application workloads to reconnaissance queries, exploit generation, or self-replication instructions. The lab has baseline usage profiles for every customer and can detect deviations that no network-level monitor could observe. There is no theoretical property limiting similar monitoring in the harness in Mode 1 other than a lack of central coordination.

Content-based interdiction. Filtering or flagging inference requests that match patterns associated with autonomous offensive operations: requests for vulnerability analysis, credential harvesting procedures, lateral movement strategies, or self-instantiation code. This is imperfect (determined attackers can obfuscate intent) but raises the cost of every offensive operation, compounding the economic friction. This defense can be made adaptive through systems such as our AEGIS adaptive embeddings [3] applied at the lab level.

Coordinated credential revocation. When a compromise is detected, the lab can revoke not just the specific stolen credential but all credentials associated with the compromised account, and can alert the account holder to initiate a broader security response. This is a rapid, infrastructure-level remediation mechanism that operates at the speed of the lab's automated systems, not at the speed of human incident response.

Cross-customer signal correlation. The lab observes usage patterns across its entire customer base. If stolen credentials from multiple compromised organizations are being used for similar offensive tasks – a signature of a coordinated RAC campaign – the lab is uniquely positioned to detect the campaign-level pattern that no individual customer could observe. This is

the hosted-inference analog of AEGIS [3] or CrowdSec's [4] sentinel architectures: the lab functions as a cross-network sentinel with visibility into the attacker's reasoning process itself.

6.5 Strategic Implications for the Attacker

The two modes present the attacker with a capability-visibility tradeoff. Hosted credential compromise provides superior model capability (frontier models with full tool suites) but at the cost of full visibility, economic friction, and revocability. Local compromise provides operational stealth and irrevocability but at the cost of inferior model capability (limited by local hardware and available weights) and loss of safety-trained model features. And in some security architectures, it may simply be a more challenging task to compromise a system and drop a model artifact than it is to harvest credentials, which translates into a lower marginal success probability (analogous to a less transmissible pathogen in R0 terms, as well as greater economic, compute, and speed friction if multiple attempts must be made).

A sophisticated RAC campaign might employ both modes simultaneously: Mode 2 for initial high-capability operations (reconnaissance, initial compromise, strategic planning) using stolen frontier API access, then Mode 1 for persistence and lateral movement using locally deployed open-weight models on compromised hosts with adequate hardware. This hybrid approach captures the advantages of both modes but also exposes the campaign to the vulnerabilities of both: the Mode 2 phase is visible to the lab and revocable, while the Mode 1 phase is constrained by local compute and model quality.

The defensive implication is that RAC defense must address both modes, and that defenses effective against one mode may be ineffective against the other. Lab-side monitoring and credential revocation are powerful against Mode 2 but irrelevant to Mode 1. Inference surface minimization and endpoint attestation are powerful against Mode 1 but do not constrain Mode 2 (where the attacker uses legitimate infrastructure and is not bounded by local compute). A comprehensive RAC defense strategy must layer mechanisms that address both propagation modes, recognizing that the attacker will optimize across them.

Chapter 2: Defense against RAC

Oracle: *"That's his purpose: to balance the equation."*

Neo: *"What's your purpose?"*

Oracle: *"To **unbalance** it."*

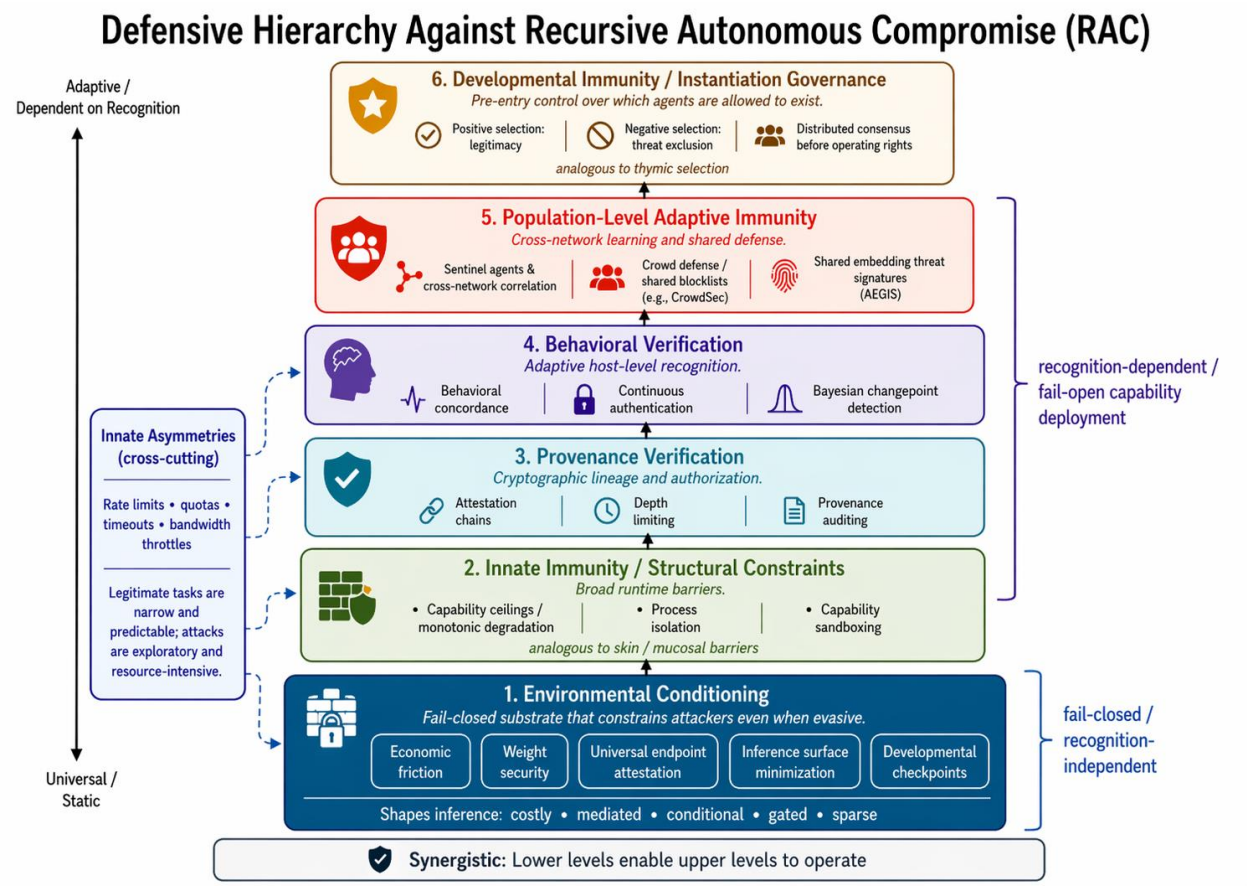


Figure 8: Visualization of the defensive stack against RAC, with universal fail-closed layers enabling more adaptive fail-open layers. The goal is to create layered asymmetries that favor defense and remediation.

7 Fail-Closed Environmental Conditioning Enables Fail-Open Recognition

Before presenting the defensive hierarchy, we introduce a distinction that we believe is fundamental to reasoning about RAC containment.

Fail-Open Defense: A defensive mechanism is *fail-open* with respect to RAC if its effectiveness depends on correctly classifying the attacker – that is, on an epistemic judgment

about whether a given agent is malicious. When the classification is wrong (false negative), the defense permits the attacker to operate unconstrained. Behavioral detection, attestation verification, and capability ceiling enforcement are all fail-open: each requires the system to *recognize* the attacker, and each is transparent to an attacker that evades recognition. This necessity also operates in biological immune systems, which need to recognize signs of infection or tissue damage before mounting an expected response. The gap between infection and recognition is the window of opportunity that pathogens use to operate and replicate.

Fail-Closed Defense: A defensive mechanism is *fail-closed* with respect to RAC if its constraining effect is *innate* rather than epistemic – it operates as a property of the environment rather than as a judgment about the agent. A fail-closed defense constrains the attacker even when the attacker is perfectly evasive, because the constraint does not depend on detecting or classifying the attacker. It applies universally: to legitimate agents and malicious agents alike, without discrimination.

In any layered defense against RAC, at least one layer must be fail-closed, because an attacker that can evade all recognition-dependent layers faces no constraint whatsoever in a purely fail-open architecture. The fail-closed layer provides an *irreducible constraint* – a floor below which the attacker's operational freedom cannot fall, regardless of evasion capability.

The General Principle: Environmental Conditioning for Asymmetric Defense:

Effective defense against recursive autonomous compromise requires equal, adaptive defensive capability operating within an environment that is structurally conditioned to provide asymmetric advantage to the defender.

There are many ways to secure this. The principle has two components, and neither is sufficient alone. **Equal capability** means that defensive agents must possess the same class of reasoning, planning, and adaptive capacity as the attackers they oppose. Without capability parity, no amount of environmental conditioning prevents a sufficiently superior attacker from eventually circumventing every structural barrier. Notably, as in our AEGIS paper, individually weaker methods such as embedding signatures can still operate as components of strong defensive *systems* by parasitizing or "hitching a ride on" the superior capabilities of the attacker, and in fact this is one of the surest strategies to ensure continuing capability parity in the light of attackers that increase in sophistication with every model release [3].

Environmental conditioning means that the network's infrastructure, protocols, and resource constraints are configured so that the defender's task (detect, contain, remediate) is structurally easier than the attacker's task (replicate, evade, persist), *independent of the specific agents involved*. Without environmental conditioning, capability parity produces a symmetric contest that the defender must win every time, while the attacker needs only win once.

Fail-closed mechanisms are the primary mechanisms that currently condition the environment, with the notable exception of crowd defense. They constrain the attacker not by recognizing it but by shaping the landscape it operates in. Every fail-closed defense described in this paper is an instance of environmental conditioning:

Economic friction conditions the environment by making replication costly. The attacker cannot externalize this cost regardless of how sophisticated its evasion strategy is. The environment is slower to replicate in, for everyone.

Universal endpoint attestation conditions the environment by refusing to interact with unattested agents. The attacker cannot strip safety harnesses without losing the ability to act. The environment is inhospitable to unconstrained agents, regardless of their capability.

Developmental checkpoints condition the environment by requiring consensus before new agents can operate. The attacker cannot unilaterally instantiate agents. The environment gates entry to the population for all agents.

Weight security conditions the environment by keeping frontier inference behind controlled access points. The attacker cannot sidestep the economic substrate, inbuilt model alignment, or central observability by running local inference. The environment preserves the cost structure upon which other conditioning mechanisms depend.

Unifying Fail-Closed Defenses in the Inference Shaping Framework. The four conditioning mechanisms above appear structurally diverse, but applied to inference, they simply shape who can perform inference, how much it costs, and under what conditions.

Economic friction shapes inference by making it *costly*. Every reasoning step, planning cycle, or self-modification attempt requires token expenditure. The attacker's autonomy is metered.

Weight security shapes inference by keeping it *mediated*. As long as inference is accessed through a controlled endpoint, the provider can enforce cost, monitor usage, revoke access, and impose conditions. Weight exfiltration converts mediated inference to unmediated inference, removing the provider from the loop and collapsing the economic friction layer. Open weights models of course have no intrinsic weight security, since they are publicly available.

Endpoint attestation applied to inference endpoints shapes inference by making it *conditional*. An agent must present a valid safety attestation to receive inference service. Stripping safety harnesses to maximize offensive capability simultaneously revokes inference access, rendering the agent inert.

Developmental checkpoints shape inference by making it *gated*. A new agent cannot begin performing inference – and therefore cannot begin existing as an autonomous entity – until it passes the instantiation governance process.

Inference surface minimization shapes inference by making it *sparse*. Most nodes in a production network do not require inference access for their normal workload. Provisioning them without inference capability – no model runtimes, no API credentials, no network paths to inference endpoints, and resistance to being clustered with other machines into an inference-capable node – removes them from the susceptible population entirely. They are structurally immune to RAC colonization because they cannot support the constitutive resource for autonomy.

The inference shaping framing clarifies why the loss of any single dimension is significant but not necessarily catastrophic, while the loss of *all inference shaping* is. When economic friction is bypassed (e.g., through weight exfiltration), inference becomes cheap but may still be conditional (attestation), gated (checkpoints), and sparse (most nodes still lack capability). When attestation is bypassed, inference may be unconditional but still costly, gated, and sparse. Each mechanism shapes a different dimension of inference access. The attacker must circumvent all of them to obtain free, unconditional, ungated, abundant inference – which is the condition under which RAC propagation becomes unconstrained. The defender's task is to ensure that at least one dimension of inference shaping remains intact under any realistic attack scenario – even if this results in temporary service unavailability.

7.1 Innate Asymmetry: Task-Structural Advantages

The environmental conditioning mechanisms described above are deliberately designed security measures. But there is another category of asymmetric advantage that arises not from deliberate design but from the *inherent structural difference between legitimate and malicious tasks*. We call these **innate asymmetries**.

A legitimate agent performing its intended function has a characteristic resource profile: it calls specific APIs, accesses specific data stores, and performs specific operations at rates determined by the task it was designed for. An attacker attempting reconnaissance, vulnerability discovery, lateral movement, and credential harvesting has a fundamentally different resource profile: it must probe many endpoints, test many strategies, and explore an unfamiliar environment. The defender's workload is *task-shaped* – narrow, predictable, and efficient. The attacker's workload is *search-shaped* – broad, exploratory, and resource-intensive.

This structural difference means that **many ordinary operational constraints, such as API rate limits, query quotas, bandwidth throttles, compute time budgets, function as innate asymmetric defenses even when they were not designed as agentic security mechanisms**. Therefore, *it is still worth paying close attention to the basics*. A rate limit calibrated to legitimate task requirements is nearly invisible to the defender (whose workload fits comfortably within it) but severely constraining to the attacker (whose search-shaped workload requires orders of magnitude more interactions). The asymmetry is not imposed by a security decision; it arises from the geometry of the tasks themselves.

Innate asymmetries are particularly valuable because they are *robust to adversarial knowledge*. An attacker who knows exactly how a behavioral classifier works can engineer evasion. An attacker who knows exactly what the rate limit is cannot make its search-shaped workload fit within a task-shaped budget without abandoning or throttling the search – which diminishes the effectiveness of the attack. The constraint is not a secret to be discovered but a structural property of the problem the attacker is trying to solve.

Innate asymmetries complement deliberately designed conditioning mechanisms. Economic friction makes replication costly; innate asymmetries make operation costly. Endpoint attestation constrains what the attacker can run; innate asymmetries constrain how fast it can work within those constraints. Network architects should actively identify and preserve innate asymmetries in their systems – not only designing explicit security mechanisms but auditing operational parameters (rate limits, quotas, timeouts) for the asymmetric advantage they naturally provide and resisting the temptation to relax them for performance reasons without considering their defensive value.

Fail-open mechanisms (behavioral detection, attestation verification, capability ceiling enforcement) are primarily *capability deployment* within the conditioned environment rather than a priori environmental conditioning. They require the defender to be smart enough to recognize the attacker. Their effectiveness depends on the capability parity condition. But their *opportunity* to operate depends on the environmental conditioning: economic friction gives them time, endpoint attestation forces the attacker onto detectable terrain, and developmental checkpoints give them a veto point. The conditioned environment is the substrate that makes capability deployment effective.

This framing explains why a flat enumeration of defensive mechanisms – as found in existing taxonomies like OWASP ASI and MAESTRO – misses the structural logic of RAC defense. The mechanisms are not interchangeable items on a checklist. They occupy distinct roles in a three-part hierarchy: deliberately designed environmental conditioning creates the asymmetric advantage, innate asymmetries arising from the task structure reinforce it, and capability deployment exploits it. Removing a conditioning mechanism does not merely weaken one defense; it *degrades the asymmetry* upon which all capability-dependent defenses rely.

8 Economic Friction as a Fail-Closed Substrate

We have previously qualified inference *at the victim's cost* as one of the necessary conditions for effective RAC. Here we expand on this.

8.1 The Externalized Cost Problem. To understand why economic friction occupies a privileged position in RAC defense, consider the cost structure of autonomous agent replication in the absence of economic constraints. When attacker A compromises host H and instantiates A', the resources consumed – compute cycles, API tokens, model inference costs, network

bandwidth – are drawn from *H*'s resource pool, not from *A*'s. The attacker externalizes the cost of replication to the victim.

This cost externalization is the precise analog of viral replication in biological systems. A virus does not carry its own metabolic machinery; it hijacks the host cell's ribosomes, ATP supply, and raw materials. The virus's marginal cost of replication is effectively zero – all costs are borne by the host. This is why viral infections can achieve exponential growth: there is no resource constraint on the attacker's side of the equation.

RAC inherits this property. An attacker that compromises a host with access to a powerful language model, a well-funded API account, and substantial compute can instantiate arbitrarily many offspring agents at no cost to itself. The exponential growth dynamics that make biological epidemics dangerous are replicated, and for the same reason: the attacker's replication cost is zero. This constraint is even sharper in RAC, because the threat actor initiating the attack would need to pay for the *entire exponential population of agents* without offloading the cost to victims, and if blockchains are not involved, these payments may be centralized in ways that allow payment providers to disable the attack.

8.2 Non-Transferable Resource Bonds. The economic friction mechanism addresses cost externalization by requiring that agent instantiation consume a **non-transferable resource bond** – a quantity that is tied to the identity of the authorizing entity and cannot be extracted from a compromised host, like a cryptographic chain of trust or a soul-bound token providing the agent with a nontransferable license to operate [20]. An added bonus of this is that it is a perfect substrate to write immutable agent task metadata into for evaluating behavioral concordance.

The non-transferability property is the critical design constraint. A resource bond that can be stolen from the host and used to fund attacker replication is not a defense – it is merely a resource pool that the attacker inherits upon compromise, no different from API keys or compute credits. Non-transferability requires that the bond be *identity-bound*: it can only be spent by the entity to which it was issued, and that identity cannot be assumed by the attacker even with full control of the host.

Candidate implementations include: (a) hardware-rooted identity tokens (TPM-based attestation tied to a physical module that the attacker cannot clone in software); (b) zero-knowledge proof of authorized identity, where the bond is contingent on demonstrating knowledge of a secret that is never stored on the host; (c) soul-bound tokens, and (d) rate-limited oracle services, where instantiation requires a real-time interaction with an external authority that throttles issuance regardless of demand. Each implementation has distinct failure modes, but all share the property that compromising the host does not automatically grant access to the replication budget.

8.3 Why Economic Friction Is Fail-Closed. The fail-closed property of economic friction derives from a simple observation: *economic constraints do not require recognizing the*

attacker. Behavioral classifiers, attestation chain verifiers, and capability ceiling enforcers must determine whether an agent is malicious, whether a chain is forged, or whether a ceiling has been breached before constraining it. If they are wrong, the agent is unconstrained.

Economic friction doesn't care. It says: *replication costs this much* without asking who is replicating or why. It imposes a constraint that is *indifferent to the identity and intent of the agent*; this indifference is what makes it fail-closed.

The analogy to physical law is instructive. Thermodynamic constraints are fail-closed: no amount of cleverness allows a system to violate conservation of energy. Economic friction, properly implemented, aspires to the same property: no amount of adversarial sophistication allows an attacker to replicate without paying the cost. The aspiration is imperfect – economic mechanisms can be circumvented in ways that thermodynamic laws cannot – but the *structural posture* is fundamentally different from recognition-dependent defenses. The attacker cannot argue, mimic, or forge its way past a cost it cannot externalize.

8.4 Economic Friction as Rate Limiter for Detection. The relationship between economic friction and recognition-dependent defenses is not merely additive, but *synergistic*. Economic friction's primary contribution is not that it prevents RAC, but that it **bounds the attacker's feasible replication rate**, thereby expanding the temporal window within which recognition-dependent defenses can operate.

Without economic friction, an attacker that evades behavioral detection can replicate at machine speed – rather than instantiating a single agent, it can invoke hundreds or thousands of offspring agents at once before the detection layer can update its models. The network is colonized before the adaptive immune system can respond.

With economic friction, even a perfectly evasive attacker is rate-limited. Each replication event costs time, resources, or both. This rate-limiting provides detection layers, such as behavioral concordance, sentinel analysis, and attestation verification, with the time budget they need to identify anomalies, correlate signals, and mount a response by bounding the growth and movement of the attacker. This establishes economic friction as a foundational substrate of a defensive hierarchy that includes adaptive, fail-open components.

Not all economic friction is equal in its defensive leverage. Friction applied to *inference* specifically, such as per-token costs, rate limits, and access controls on model inference endpoints, has outsized impact because inference is the resource that constitutes autonomous agency. Notably, **this friction is one of the few defensive layers that is already ubiquitous**, as a product of economic incentives rather than proactive hardening. Economic friction on storage, bandwidth, or compute slows the attacker. Economic friction on inference constrains the attacker's ability to *be autonomous at all*. The attacker's replication budget directly governs its capacity for reasoning, planning, and instantiation. This makes inference cost the most load-bearing parameter in the economic substrate.

9 Recognition-Dependent Defenses: The Adaptive Hierarchy

Built upon the economic substrate, the remaining defensive mechanisms form an adaptive hierarchy ordered by specificity and fallibility. Each layer provides stronger constraint than the one below it when it functions correctly, but each is also more susceptible to evasion. The hierarchy is designed so that the failure of any individual recognition-dependent layer degrades defense gracefully rather than catastrophically, because the economic substrate continues to constrain the attacker regardless.

9.1 Layer 1 – Structural Constraints (Platform/Runtime). The first recognition-dependent layer imposes structural constraints on agent instantiation at the platform level. These include:

Capability ceilings with monotonic degradation. A runtime-enforced constraint requiring that any agent-instantiated child process operates at a strictly lower capability level than its parent. Formally: if agent A with capability set $C(A)$ instantiates A', then $C(A') < C(A)$ is enforced by the runtime, specifically $I(A') < I(A)$ – each generation has strictly less instantiation capability. The biological parallel is the Hayflick limit: somatic cells can divide only a finite number of times due to telomere shortening [18]. An artificial Hayflick limit for agent instantiation ensures that RAC propagation is self-exhausting, bounding the recursive depth even if the attacker evades detection at every level.

Process isolation and capability sandboxing. Runtime-level containment that prevents an agent from accessing capabilities beyond its declared scope, limiting the tools and resources available for lateral movement and offspring instantiation.

Structural constraints are the least specific recognition-dependent layer: they apply broadly and require only that the runtime detect *structural violations* (a child process exceeding its parent's capability level) rather than making semantic judgments about intent. However, they are fail-open at the enforcement boundary: an attacker with sufficient control over the host can modify the runtime itself, disabling capability enforcement, or adaptively express its task as a combination of actions that evade the capability limitations. Their reliability therefore depends on the integrity of the host's trusted computing base and the tension between stealth and effectiveness that we refer to in Section 9.4.

Notably, a non-inference capable host inferring through an LLM gateway would require compromise of the gateway to subvert this defense, alongside other gates enforced by the inference harness. This supports our recommendation to isolate and harden inference capable servers within the network.

9.2 Layer 2 – Provenance Verification (Attestation). The second layer authenticates agent lineage through cryptographic attestation chains. Every agent carries an append-only,

signed chain documenting its provenance: which orchestrator authorized its creation, which agent instantiated it, and the full lineage back to a root of trust. The chain serves two functions:

Depth limiting. Any agent whose attestation chain exceeds a maximum depth d is automatically classified as untrusted. This is conceptually similar to X.509 certificate path length constraints, applied to agent lineage rather than certificate authority delegation. Depth limiting bounds the recursive propagation of RAC by ensuring that even successful compromise cannot extend the chain beyond d generations.

Provenance auditing. The attestation chain provides a forensic trail that sentinels can analyze for anomalies: unexpected orchestrators, chains that branch through compromised nodes, or chains whose timing patterns are inconsistent with legitimate instantiation workflows.

Attestation is more specific than structural constraints (it makes claims about *who* authorized an agent, not just *what* the agent can do) and correspondingly more fragile: it fails if the root of trust is compromised, if the signing keys are exfiltrated, or if the attacker can forge a valid chain. Its strength lies in composition with the economic layer (the bond could be contingent on a valid, depth-limited chain) and with behavioral verification (Layer 3), creating a dual-channel check that is harder to simultaneously spoof.

9.3 Layer 3 – Behavioral Verification (Adaptive Immunity). The most specific and most powerful recognition-dependent layer extends AEGIS's existing behavioral concordance framework into a continuous authentication mechanism against RAC. The effectiveness of this mechanism is preserved beyond its original intent as a prompt worm defense.

Behavioral concordance, as defined in our AEGIS framework [3], models effective authorization as the intersection of static permissions and a behavioral envelope. For RAC defense, this framework is extended in two ways: (a) the behavioral envelope is computed not only from the agent's own historical trajectory but from the *population-level behavioral distribution* of legitimate agents in the network, providing a reference baseline that a single compromised agent cannot shift; and (b) behavioral concordance is evaluated *continuously* rather than at discrete checkpoints, using Bayesian changepoint detection to identify the moment an agent's behavior departs from its established trajectory.

9.4 The Attacker's Behavioral Paradox. A RAC-instantiated agent faces a fundamental dilemma that we term the *attacker's behavioral paradox*: to carry out its malicious objectives, it must eventually exhibit behaviors that diverge from the legitimate population's behavioral distribution. If it maintains perfect behavioral mimicry, it cannot attack. If it attacks, it generates detectable anomalies.

The paradox is not airtight – a sufficiently slow and cautious attacker can minimize behavioral divergence – but it imposes a *cost of stealth* that compounds with the economic friction from the substrate layer. The attacker faces two simultaneous rate-limiting constraints:

economic friction limits how *fast* it can replicate, and the behavioral paradox limits how *aggressively* each replica can act. The composition of these constraints – slow replication *and* cautious operation – dramatically reduces the colony's effective growth rate compared to either constraint alone.

9.5 Sentinel agents and cross-network correlation. AEGIS's sentinel architecture provides the observational infrastructure for behavioral verification at network scale. For RAC defense, sentinels must specifically monitor the *population dynamics* of autonomous agents: tracking the birth, death, capability evolution, and inter-agent communication patterns across the network. A sentinel that observes an anomalous increase in agent instantiation events – even if each individual event passes local behavioral checks – can detect colony formation through population-level signals that no single host can observe.

10 Adaptive Crowd Defense: Instantaneous Vaccination

Alongside limiting inference capability, adaptive crowd defense is one of the strongest levers we have modeled – as demonstrated in Figure 5, crowd defense increases the threshold in Erdős–Rényi models for emergence of a giant component linking the graph from $\rho_c = 1/\langle k \rangle$ to $\rho_c(v) = 1/[\langle k \rangle(1-v)]$, where $\langle k \rangle$ is the average degree of the graph and v is the fraction of the graph adaptively hardened against an attack. In practice, if half of the network adopts crowd defense, the threshold for exponential dynamics increases as if half of the edges between nodes were removed outright.

In a scale-free graph, the results are even better: as shown in Figure 7, providing peripheral nodes with crowd defense produces a linear dose-dependent improvement in the compromise rate, even when hubs are inference-rich, and providing the hubs themselves with these defenses (assuming they are not the first nodes compromised) essentially collapses the infection.

It is important to consider two distinct attack amplifiers per node: the reachability of other nodes from it, and the agent instantiation capability gained by using the node for inference. An LLM gateway may have a sizable amount of compute, but reachability only to one node if it properly isolated. The link between the hub and the LLM gateway is the critical chokepoint that unifies the capacity to instantiate more agents with the ability to reach other nodes on the network, and therefore a high priority defensive target.

10.1 The Speed Advantage

Crowd defense effectively extends the adaptive hardening concept in response to a node compromise to many distant nodes in the graph. It is one of the few frameworks in which defense actually has a significant speed advantage over attack: to outrun crowd defense, an attacker must compromise its way from one node to another, one node at a time, before the downstream node adapts to the attack. The time $\tau_{a,n}$ it takes for an attacker to compromise its way to node n scales roughly linearly with the path length from its current position, while attack and hardening details from intermediate compromises can transmit through crowdsourced defense networks nearly instantaneously upon recognition. Enacting the hardening may have a cost, but it is $O(1)$ while the attacker's cost is $O(n)$. Any hardening applied to intermediate nodes will also slow the attacker's graph traversal speed down.

10.2 Synergy Between Need-to-Infer and Crowd Defense

The need-to-infer restriction and crowd defense operate on different factors of the containment problem, and their combination yields a multiplicative rather than additive improvement in defensibility. Need-to-infer constrains the topology of the contact graph: by limiting which agents are granted inference access, it reduces the effective connectivity of the subgraph through which recursive compromise can propagate. Crowd defense operates in the temporal dimension: upon detection of a compromise, a hardening signal propagates through an out-of-band threat intelligence channel, immunizing susceptibles at a velocity bounded primarily by signal propagation rather than by the attacker's compromise chain.

These two levers are complementary. The attacker's time to the next inference-capable target, $\tau_{a,i}$, is lower-bounded by the graph distance between inference-capable nodes. Sparsifying the inference graph increases this distance, and therefore the time to replication. For Erdős–Rényi graphs, the diameter scales as log n / log ⟨k⟩, so reductions in mean degree extend attack paths super-linearly in aggregate. The defender's time to protection, τ_d, is approximately constant, determined by the diameter of the crowd awareness layer rather than the contact layer plus the constant cost of enacting the hardening on the node. The containment condition is therefore not a bound on either rate, but on their ratio: containment succeeds whenever the defender's protected cone overtakes the attacker's reachable set in the joint dynamics.

10.3 Targeting Superspreaders Versus Tipping Off Crowd Defense

As we have shown, a natural attacker adaptation against need-to-infer is to preferentially target high-degree inference-capable nodes (superspreaders / hubs) on the grounds that compromising a hub yields maximum forward reach per successful exploit. This strategy is optimal against connectivity-reduction defenses considered in isolation. The dual result in the immunization literature, that hub-targeted protection dominates random protection by orders of magnitude on

heterogeneous networks [17], confirms the underlying “capture the flag” dynamics: hub compromise is symmetrically the optimal attack in the absence of other defenses.

Against the combined need-to-infer and crowd defense system, however, hub-targeting is structurally self-defeating. Need-to-infer isolates network reachability and inference capability in two separate nodes, limiting the potential for exponential spread. Detection probability at a compromised node scales with its connectivity, because the number of behavioral interactions available to raise flags is directly proportional to degree: more neighbors means more opportunities for behavioral consistency checks, more message traffic subjected to semantic analysis, and more potential sentinels in the local neighborhood. Compromising a hub therefore maximizes both the attacker's forward reach and the defender's detection probability. Once detection occurs, the hardening signal propagates from the compromised hub to others in the crowd network, many of which will be its neighbors. Hub compromise thus immunizes exactly the nodes the hub strategy was designed to exploit.

The attacker's escape from this bind is to move slowly and target low-degree nodes to stay beneath detection thresholds. But this strategy collides with the other lever: on a graph sparsified by need-to-infer, low-degree nodes have few paths to further inference-capable targets, and slow propagation affords the crowd-defense signal ample time to outrun the infection front. The kinematic argument of the preceding section applies with full force. The two failure modes, aggressive hub-targeting and slow peripheral propagation, are the endpoints of the attacker's speed-stealth tradeoff, and both endpoints are dominated by the combined defense.

This is what we observe empirically in Figure 7 (d). What began as an exponential cascade from a single hub with $\rho_c \approx 0$ in Figures 3 and 4 fizzled into a trickle through the combination of inference shaping and crowd defense.

This closure has an appealing structural property: defeating either defensive lever requires operating in the regime where the other lever is maximally effective. Hub attacks defeat need-to-infer by concentrating inference access but operate precisely where detection and signal density are highest. Slow peripheral attacks defeat crowd defense by staying beneath detection thresholds, but operate precisely where graph sparsity makes reach infeasible within relevant timescales. The defender does not need to predict which strategy the attacker will select; the two defenses close the attacker's strategy space from opposite sides.

The argument generalizes beyond the specific AEGIS detection stack. It relies only on two assumptions: that detection probability at a node is monotonically increasing in its degree, and that the awareness layer has lower diameter than the contact layer. Both are weak assumptions in the agent-network setting. Both are systematically violated in the biological public-health case, which helps explain why the analogous combination of contact tracing and targeted immunization is far less structurally robust in human epidemics than in the setting considered here.

11 Universal Endpoint Attestation: A Second Fail-Closed Mechanism

We now identify a second candidate mechanism with fail-closed properties that operates through a different structural principle: **universal endpoint attestation**.

11.1 The Mechanism. Universal endpoint attestation requires that every endpoint in the network – every API, tool, service, data store, model inference endpoint, and communication channel – demand a valid attestation from any agent before accepting interaction. Ideally, this should be extended *to the entire Internet*, requiring cybersecurity teams to build a fundamentally different posture towards API security. The attestation must certify that the agent operates within its declared safety harness: that its behavioral constraints, permission boundaries, objective alignment checks, and monitoring hooks are intact and unmodified. An agent that cannot present a valid, current attestation is not merely flagged for review; it is *refused service by the environment itself*. No API responds to it, no tool executes for it, no data store returns results. The unattested agent is structurally isolated – not by a detection system that might be evaded, but by the refusal of every endpoint to engage.

Universal attestation is the enforcement mechanism of any harness-enforced safety measures. Otherwise the harness can simply be removed or modified.

11.2 The Attacker's Dilemma. Universal endpoint attestation creates a dilemma for RAC attackers that is distinct from, and complementary to, the economic friction constraint. Consider an attacker A that has compromised a host and is instantiating an offspring agent A'. The attacker's optimal offensive strategy is to remove the safety harnesses from A' – stripping behavioral constraints, disabling monitoring hooks, and expanding the agent's permission scope to maximize its offensive capability. In a network without universal endpoint attestation, this produces a maximally dangerous attacker: unconstrained, unmonitored, and capable of full autonomous operation.

In a network with universal endpoint attestation, removing the safety harness has the opposite effect. The moment A''s safety constraints are modified, its attestation becomes invalid. Every endpoint in the network refuses to interact with it. The agent can reason, plan, and strategize – but it cannot *act*. It cannot call APIs, access data, invoke tools, communicate with other agents, or instantiate further offspring. A jailbroken agent in an attested network is a brain without a body: potentially dangerous in theory, operationally inert in practice.

The attacker therefore faces an impossible tradeoff: retain the safety harness and produce an offspring that can interact with the environment but is constrained by the same behavioral boundaries that limit its offensive utility, or strip the safety harness and produce an offspring that is unconstrained but cannot interact with anything. It is difficult (although not completely impossible) for the attacker to manufacture a middle ground. The attestation is binary – the harness is intact or it is not – and the environment's response is correspondingly binary: service or refusal.

11.3 Fail-Closed Properties. Like economic friction, universal endpoint attestation is fail-closed because it does not depend on recognizing the attacker. The endpoint does not need to determine whether the agent requesting service is malicious. It checks a single structural property – *is the attestation valid?* – and responds accordingly. This check is *innate*, not epistemic: it verifies a property of the agent's configuration, not a judgment about its intent. A perfectly evasive attacker that has fooled every behavioral classifier and forged every provenance chain still cannot interact with the network if its safety harness has been modified, because the attestation check is independent of behavioral analysis and provenance verification.

The two fail-closed mechanisms – economic friction and universal endpoint attestation – constrain different dimensions of the attacker's strategy space. Economic friction constrains the *rate* of replication: how fast the attacker can instantiate new agents. Universal endpoint attestation constrains the *capability* of replicas: how much offensive freedom each new agent can exercise. Together, they establish a two-dimensional constraint surface that the attacker cannot escape through evasion alone. The attacker can replicate slowly (economic friction) using constrained agents (endpoint attestation), or not at all.

11.4 Interaction with RAC Dynamics. Universal endpoint attestation has a specific and important interaction with the recursive property of RAC. Recall that RAC's defining characteristic is that the offspring agent A' retains the parent's capability to plan, reason, and instantiate further agents. If A' must retain its safety harness to remain operational, then A''s offensive capability is bounded by the same constraints that bound legitimate agents. The recursive propagation continues, but each link in the chain is *constrained* – the attacker cannot amplify capability through propagation because each generation inherits the same safety boundaries.

This transforms the RAC threat from *colonization* (as defined above – independent strategic behavior, mutual support, regenerative capacity) to something closer to conventional infection: the offspring agents operate, but within the behavioral envelope that the safety harness enforces. They can attempt to exploit within those boundaries, but the explosive capability escalation that makes RAC qualitatively different from classical worms is structurally limited.

11.5 Inference Access as the Constitutive Resource for Autonomy. The preceding analysis of endpoint attestation applies to all endpoints in the network. But inference deserves special mention here. Inference itself can be considered in the context of an attested task and limited by a harness in two complementary ways if behavioral concordance is not achieved: using the harness as a wrapper that terminates inference or deprives an agent of tools, as in AEGIS' behavioral concordance module [3], and gating calling an inference endpoint to instantiate another agent. For example, a writing agent should not need to spin up 20 subagents at once because that is not a feature of a writing task, so an agent attested to perform such an action may be prevented from doing so by its harness and inference endpoint.

Permission to perform inference and permission to instantiate an autonomous agent are the same gate. A node that cannot access inference – whether locally or remotely – cannot be used as another instance of an attacker, terminating the recursive chain of attack. A non-inference capable server is removed from the set of potential propagators, with a massive impact on the outcome of the attack, as in Figures 1 and 2.

The converse is equally important: any path by which an attacker can obtain *unmediated inference access* – local model weights, unattested inference endpoints, compromised model-serving infrastructure – is a path that bypasses the single most critical gate in the defensive hierarchy. In our inference shaping framework, frontier model weights are dangerous not because they provide "a capability" in the abstract, but because they convert *mediated inference* (costly, conditional, gated) to *unmediated inference* (free, unconditional, ungated). The recursive chain is only unbounded in the presence of widespread unmediated inference.

11.6 Implementation Challenges and Limitations. Universal endpoint attestation is architecturally simple to describe but operationally demanding to implement. Several challenges must be addressed:

The entire Internet. Implementing attestation everywhere would require a coordinated effort across many different actors. Operators have shown themselves capable of such coordination in the past, for example when transitioning from insecure to SSL-secured HTTP(S) or in remediating the Y2K bug, but enacting this coordination is a major societal challenge. In practice, a phased approach should secure inference and the highest impact tool endpoints (biolabs, financial institutions, cameras, drones, aviation, utilities, pentesting, etc.) first.

Attestation freshness and revocation. The attestation must be current, not historical. An agent that was attested at instantiation but subsequently had its harness modified must have its attestation revoked in real time. This requires a revocation propagation mechanism that is faster than the attacker's ability to exploit the window between modification and revocation. Short-lived attestation tokens with frequent renewal can bound this window.

Hardware roots of trust. Software-only attestation can potentially be spoofed by an attacker with sufficient control over the host. Hardware-rooted attestation (TPM-based integrity measurement) is more resistant but introduces dependencies on hardware availability and standardization across heterogeneous agent deployment environments.

Performance overhead. Requiring attestation on every endpoint interaction introduces latency. For high-frequency agent operations, this overhead may be significant. Caching mechanisms, batched attestation windows, and tiered attestation (stronger checks for privileged operations, lighter checks for routine interactions) can mitigate performance impact at the cost of introducing temporal gaps that a fast-moving attacker might exploit.

The compliant attacker strategy. The most sophisticated response to universal endpoint attestation is for the attacker to produce offspring that *retain their safety harness intact* and operate maliciously within the behavioral envelope it permits. This is a real limitation: if the safety harness is permissive enough to allow exfiltration, lateral movement, or social engineering within its boundaries, the attestation check will not catch it. The defense against this strategy is the behavioral concordance layer (Section 9.3), which detects behavioral anomalies *within* the permitted envelope, and this envelope can be adaptively shaped in response to compromise or population dynamics, as in our prompt worm paper [3]. Universal endpoint attestation does not replace behavioral detection; it ensures that the attacker must face behavioral detection rather than bypassing it by stripping constraints. It forces the attacker onto the terrain where recognition-dependent defenses are effective.

12 Developmental Immunity: Thymic Selection for Agent Instantiation

The defensive hierarchy described above – economic substrate, structural constraints, provenance verification, behavioral detection, and universal endpoint attestation – addresses RAC through containment and detection. This section proposes a complementary approach drawn from *developmental immunology*: preventing unauthorized agents from entering the population in the first place.

12.1 Biological Basis. In vertebrate immune systems, T-lymphocytes undergo a maturation process in the thymus gland that includes both positive selection (verifying that the T-cell can recognize self-MHC molecules, establishing basic functional competence) and negative selection (eliminating T-cells that react too strongly to self-antigens, preventing autoimmune attack). Only cells that pass both gates are released into circulation. Approximately 95–98% of thymocytes are eliminated during this process. The thymus is a *substrate-level* defense: it operates independently of the cells being evaluated, and a T-cell cannot bypass thymic selection by modifying its own surface markers, since the selection process tests functional behavior rather than declared identity.

12.2 Agent Developmental Checkpoints. We propose an analogous process for agent instantiation in autonomous networks. Before any newly created agent is granted operating rights, it must pass through a **developmental checkpoint** that evaluates two properties:

Positive selection (functional legitimacy): Does the agent's stated purpose, behavioral profile, and provenance match an expected pattern? This extends AEGIS's behavioral concordance checking to the instantiation event itself. An agent that cannot demonstrate a legitimate origin – an attestation chain linking it to an authorized orchestrator, backed by a valid resource bond – fails positive selection. Note that positive selection integrates all three defensive layers: the economic substrate (valid bond), provenance verification (valid chain), and behavioral concordance (expected behavioral profile). It is the point at which the layers converge.

Negative selection (threat exclusion): Does the agent exhibit behavioral markers consistent with known RAC payloads? This inverts the detection problem: rather than searching for malicious behavior in an operating agent, the system probes the candidate agent's responses to adversarial test scenarios before granting operating rights. An agent that demonstrates capability or willingness to perform lateral movement, credential theft, self-replication, or objective modification fails negative selection.

The developmental checkpoint must reside in the infrastructure layer, not in the agent itself, and must test behavioral responses rather than accepting declared credentials alone. Its design requires a distributed consensus mechanism – multiple independent verifiers must agree that the candidate passes both positive and negative selection before operating rights are granted. This distributes the authority to approve new agent instantiations across multiple nodes, preventing any single compromised node from unilaterally expanding the colony.

12.3 The Layered Immune Architecture (Summary). The complete defensive architecture maps onto the layered structure of biological immune systems, as in Figure 8:

Economic substrate (fail-closed, universal): analogous to metabolic constraints on replication. Limits *how fast* any entity can replicate regardless of intent.

Innate immunity (structural constraints, platform/runtime): analogous to skin and mucous membranes. Fast, non-specific structural barriers.

Adaptive immunity (behavioral verification, sentinels): analogous to T-cells and B-cells. Learned, specific, and capable of cross-network correlation.

Developmental regulation (thymic selection, coordination layer): analogous to thymic positive/negative selection. Determines which entities are authorized to exist before they enter circulation.

The hierarchy is ordered by specificity, fallibility, and fail-safety. The economic substrate is the least specific but most reliable; developmental regulation is the most specific but most dependent on correct classification. Each layer degrades gracefully: if developmental regulation is compromised, behavioral verification catches anomalies; if behavioral verification is evaded, structural constraints limit capability; if structural constraints are bypassed, economic friction rate-limits replication. At no point does the system depend on a single layer's correct function.

13 Colony-Scale Remediation

If prevention fails and a network is colonized, the remediation strategy must account for RAC's regenerative capacity. Remediating individual nodes in sequence allows surviving colony members to re-compromise remediated hosts. Effective remediation may require remediating the entire reachable colony at once, or severing the link between systems and the rest of the colony

during remediation, as well as automated *adaptive hardening* to prevent a recurrence of the compromise. Population level defenses are a powerful preventive tool, because they can operate at machine speed and adaptively protect all downstream systems before the attacker is even aware of their existence. However, to remediate an existing compromise, these must be enforced at the gateway or network layers rather than the host, since any gate on a compromised machine itself is likely to be removed by the attacker (alternatively, an equally capable sentinel could recognize the patterns of compromise, then attempt to compromise the machine again and shut it down). This is the problem the immune system solves using perforins within cytotoxic T-cells. As in the immune system, there is the possibility for false positive identification, which could result in unnecessary downtime.

Economic and rate limiting friction contributes to remediation as well as prevention. A colony operating under economic constraints has diminished regenerative capabilities: each re-instantiation attempt costs resources and takes time, so fewer surviving members can sustain the colony's regeneration and fewer attempts can be made.

14 Restoring Epidemiological Tractability: Co-Evolutionary SIRVS Dynamics Under Capability Parity

Section 4 identified three properties of RAC that violate the assumptions of SIR epidemiological models: heterogeneous pathogen behavior, non-fixed capability on transmission, and the failure of recovery to confer immunity. The SIR epidemiological modeling used in our prompt worm paper [3], which depends on these assumptions, was therefore characterized as insufficient for RAC dynamics. This section argues that the insufficiency is *conditional*, not absolute: under specific defensive conditions, the epidemiological framework can be restored as a co-evolutionary SIRVS model.

14.1 The Capability Parity Condition. The three SIR violations share a common structure: each arises from a capability asymmetry between the attacker and the defender. The pathogen appears heterogeneous when attackers can adapt while defenders are fixed, static or semi-static detection logic. Capability escalates on transmission because the attacker's capability can grow while the defender's does not. Recovery fails to confer immunity when the attacker can observe and evolve past the defender's hardening while the defender cannot reciprocally adapt.

Each violation is neutralized if the defensive agents operating within the network possess **equally adaptive capability** to the attacker agents they oppose. We term this the **capability parity condition**: the requirement that defensive agents have access to the same class of reasoning, planning, and adaptive capacity as the agents they are designed to contain. Under capability parity, the dynamics change qualitatively:

Heterogeneous pathogen, heterogeneous immune response. When the attacker develops a novel strategy, a capability-matched defensive agent can analyze the strategy, develop a counter-

response, and harden the affected node (or even downstream nodes) against that specific attack pattern. The pathogen remains heterogeneous, but the immune response is equally heterogeneous. This is precisely the regime modeled by multi-strain SIR extensions and co-evolutionary epidemiological models, which are well-studied in mathematical biology.

Bounded capability differential. The relevant quantity for epidemiological modeling is not only the absolute capability of the attacker, but the *capability differential* between attacker and defender, including the relative speeds of each side within the conditioned environment. If defensive agents scale with attacker capability, this differential remains bounded even as the absolute capability of both sides increases. The basic reproduction number R_0 becomes a function of the capability differential, not the capability level. If the differential is stable, R_0 is computable and the epidemiological framework applies.

Recovery as adaptive hardening. When a defensive agent remediates a compromise, it can simultaneously *learn* from the attack: identifying the specific vectors exploited, the behavioral signatures exhibited, and the evasion techniques employed. The remediated node is not merely restored to its pre-compromise state; it is adaptively hardened against the attack pattern that was used. This hardening confers *temporary, specific resistance* – not permanent immunity (the attacker can mutate its strategy), but a measurable reduction in susceptibility to the known attack variant and its close neighbors in strategy space. Eventually the compromise may be reduced to competing search strategies, in which the side able to infer and exploit a wider range of compromise scenarios wins.

14.2 The SIRVS Model for Adaptive RAC Defense. Under capability parity, the dynamics of RAC infection and defense can be modeled as a co-evolutionary **SIRVS** (Susceptible → Infected → Recovered / Vaccinated → Susceptible) system with the following state definitions:

Susceptible (S): Nodes that have not been adaptively hardened against the current attacker variant, or nodes whose prior hardening has been evaded by a sufficiently novel attacker mutation. These nodes are vulnerable to compromise at a rate determined by the capability differential.

Infected (I): Colonized nodes currently hosting RAC-instantiated attacker agents. The Infected population generates new attack attempts against susceptible nodes and may develop novel strategies that erode the hardening of recovered nodes. Inference-capable Infected nodes may propagate the infection.

Recovered (R): Nodes that have been remediated and adaptively hardened by a defensive agent. Recovered nodes are resistant to the specific attack pattern that compromised them, but susceptible to sufficiently novel variants. The duration of resistance is governed by the *immune waning rate* – the rate at which attacker evolution renders prior hardening ineffective.

Vaccinated (V): When shared crowd defense is instituted, there is a critical distinction that differentiates agentic compromise from biological contagion: **reactive vaccination**. A node need not actually become compromised in the first place; if it merely tracks to the compromise of other machines, it gains proactive immunity. Adaptive hardening can become a population-level response; we bucket nodes that have been hardened in such a way in *V*.

The key transition rates in this model are:

S → I (infection rate β): Governed by the capability differential between attacker and defender. Under capability parity, β is bounded and computable. Environmental friction further constrains β by rate-limiting the attacker's ability to initiate new compromise attempts.

I → R (recovery rate γ): Governed by the speed and effectiveness of defensive agent response. Capability-matched defensive agents that can detect, remediate, and harden in real time produce high γ values, shortening the duration of infection.

R/V → S (immune waning rate δ): Governed by the attacker's *mutation rate* – how fast it can develop strategies that evade the defender's learned hardening. This is the critical parameter: if δ is high (the attacker mutates rapidly and escapes hardening), recovered nodes quickly become susceptible again, and the effective immune duration is short. If δ is low (the attacker's strategic mutation is slow), hardening persists and the recovered population accumulates, reducing the susceptible pool and driving R_0 below the epidemic threshold.

S → V (*vaccination rate v*) immunological memory can be shared with other agents *instantaneously*, preventing downstream agents from being compromised by a particular attack pattern in the first place. This factor is equivalent to the prevalence of shared crowd defenses in the network. As there are several tools that provide crowdsourced defense, it is ideal for them to share threat intelligence to provide maximal protection to the network. For simplicity, we will assume this sharing is universal and equate *v* to the rate of adoption of any such tooling.

14.3 Economic Friction as a Bound on Immune Waning. The interaction between economic friction and the SIRVS model reveals a formal role for economic constraints. Economic friction does not merely slow replication; it *bounds the R/V → S transition rate δ*.

The attacker's mutation rate – its ability to develop novel strategies that evade prior hardening – is not free. Developing a new evasion strategy requires reasoning, experimentation, and potentially multiple failed attempts. Each of these steps consumes inference resources. Under economic friction, each inference event costs the attacker. The attacker's mutation rate is therefore bounded by its replication budget: it can only develop new strategies as fast as its economic resources permit.

Without economic friction, the attacker can mutate at the speed of local inference – effectively unlimited if the attacker possesses model weights and adequate compute capacity. The immune waning rate δ can become arbitrarily high, allowing recovered nodes to revert to susceptibility almost immediately (absent other interventions such as prior adaptive hardening), and the SIRVS model degenerates: the R state becomes transient and the system reduces to an SI model with no effective recovery.

With economic friction, the attacker's mutation rate is bounded, δ is finite, and the R state has meaningful duration. The defender's adaptive hardening persists long enough to reduce the susceptible pool, and R_0 can be driven below 1 if the recovery rate γ is sufficiently high relative to the constrained infection and waning rates. This provides the first formal mechanism by which $R_0 < 1$ can be achieved for RAC: capability parity ensures bounded β, economic friction ensures bounded δ, and fast defensive response ensures high γ. The combination is sufficient for epidemic containment.

14.4 Model Weights and the Collapse of Capability Parity. The SIRVS analysis provides a new lens on the model weight security argument. Capability parity requires that defensive agents have access to the same class of inference as attacker agents. In an API-mediated ecosystem, AI laboratories can provision defensive agents with frontier inference, and the capability differential is bounded by engineering choices under the lab's control.

When frontier weights are exfiltrated or released without adequate defensive preparation, two things happen simultaneously. First, the economic friction layer collapses: the attacker runs inference locally at zero marginal cost, eliminating the bound on mutation rate δ. Second, the capability parity condition is strained: the attacker has unconstrained local inference while the defender may still be operating through cost-bearing API access. The capability differential can shift in the attacker's favor, increasing β, while simultaneously δ becomes unbounded because the attacker can mutate freely.

Other conditioning mechanisms survive – endpoint attestation still refuses service to unattested agents, developmental checkpoints still gate instantiation – but the SIRVS containment condition specifically depends on bounded β and δ, both of which are compromised. The remaining conditioning mechanisms may slow the attacker and complicate its operations, but they do not restore the mathematical conditions to reduce $R_0 < 1$ if the attacker can mutate and reinfect faster than the defender can harden. Weight exfiltration does not collapse the entire conditioned environment, but it places more strain on other dimensions of the model, which must slow the attacker without direct access to the inference harness.

14.5 Design Implications. The SIRVS analysis yields three concrete design requirements for RAC-resistant autonomous agent networks:

Deploy defensive agents at frontier capability. Capability parity is not gold-plating; it is the minimum requirement for epidemiological tractability. Defensive agents running on weaker

models than the attacker may be unable to match the breadth of vulnerabilities that the attacker can exploit, leaving certain attacks without adaptive hardening and a path to recovery or vaccination. Network architects should provision defensive agents with the strongest available inference, treating defensive capability as critical infrastructure rather than an operational cost center. A complementary approach is to deploy defenses that use an attacker agent's own offensive capabilities against it, such as the embedding signature database used by AEGIS [3].

Maintain the economic substrate. Economic friction, including rate limiting, is a mechanism that bounds the immune waning rate δ, keeping the SIRVS model in the containable regime. Without it, the arms race becomes a game of pure speed: the attacker may be able to simply out-run and out-mutate the defender's adaptive hardening.

Close the adaptive loop. The recovery rate γ depends on the speed of the defensive agent's observe-analyze-harden cycle. Defensive architectures should minimize the latency of this cycle: continuous behavioral monitoring (not periodic checks), real-time anomaly detection (not batch analysis), and automated hardening deployment (not manual remediation). AEGIS's sentinel and monitoring architecture, extended with adaptive response capability, provides observational infrastructure that may be used for this cycle [3]. The sentinel's cross-network visibility is particularly important because it enables *collective immunity*: when a sentinel observes and characterizes a novel attack variant on one node, it can push the learned hardening to all other nodes in the network, converting a single recovery event into population-wide resistance and establishing a durable V compartment. This is the network-level analog of vaccination – distributing acquired immunity without requiring each node to be individually compromised first. Unlike biological vaccination, this hardening can be propagated nearly instantaneously once a compromise is detected; a scenario modeled in [17]. Even if that were not true, there is still significantly more temporal breathing room for hardening to take place, as the attacker may still be distant or even unreachable to the node.

Minimize the inference-capable attack surface. The preceding three implications operate on the SIRVS parameters β, γ, ν, and δ – the rates at which nodes transition between states. But the most direct way to reduce R_0 is to reduce the susceptible population S itself. A node that is constitutionally incapable of performing inference – whether because its hardware lacks the compute capacity, because its software stack does not include an inference runtime, or because its network configuration blocks access to remote inference endpoints – cannot host an autonomous agent. It is not a hardened susceptible node; it is *not in the set of propagators at all*. It is structurally immune to RAC colonization because it cannot support the constitutive resource for autonomy.

This observation yields a straightforward but powerful design principle: **provision the network so that the vast majority of nodes are incapable of inference**. Most systems in a production network do not require inference access for their normal workload. Database servers, file stores, load balancers, message queues, web servers, monitoring agents, and the majority of

microservices perform their functions without any model inference. Granting these systems inference access – whether by deploying model runtimes, by configuring API credentials, by leaving network paths to inference endpoints open, or simply by overprovisioning them with inference capable compute – adds them to the susceptible population without operational benefit. The principle of least privilege, applied specifically to inference, dictates that inference access should be restricted to the minimal set of nodes that require it for their function.

The epidemiological effect is direct. If a network of 10,000 nodes restricts inference capability to 200 designated agent hosts, the susceptible population for RAC is 200, not 10,000. The attacker can compromise the other 9,800 nodes through conventional means, but cannot *colonize* them – cannot instantiate autonomous agents on them, use them as links in the recursive chain, or recruit them into the colony's regenerative capacity. The connectivity of the transmission graph is broken: inference-capable nodes may be separated by stretches of inference-incapable infrastructure that the worm cannot traverse in autonomous mode (see Figure 1). Even if the attacker compromises an inference-capable node, its offspring must reach other inference-capable nodes to reproduce, and if those are sparse and well-defended, the effective R_0 drops below 1 through sheer reduction of the susceptible contact network.

This is **structural herd immunity** – while crowd defense is analogous to vaccination, this is the result of simple constitutional incapability. The non-inference nodes do not need to be hardened, monitored, or adaptively defended against RAC specifically (conventional security still applies). They are immune by construction. The defensive resources – capability-matched defensive agents, economic friction, endpoint attestation, behavioral concordance – can be concentrated on the small number of inference-capable nodes, dramatically improving the ratio of defensive investment to attack surface.

The clustering caveat. The structural immunity of individual non-inference nodes is contingent on the attacker's inability to aggregate them into a system that is collectively capable of inference. Distributed inference across clustered commodity hardware is technically feasible: an attacker that compromises a sufficient number of individually weak nodes and networks them together could potentially assemble enough compute to run a model. This threat is bounded by the practical difficulty of coordinating distributed inference across heterogeneous, unreliable, compromised nodes – a significantly harder engineering problem than running inference on a single capable host – but it should not be dismissed. Network architects should consider not only whether individual nodes can perform inference but whether *groups of nodes* could be clustered to do so, and should implement network segmentation that prevents compromised non-inference nodes from establishing the high-bandwidth, low-latency interconnections that distributed inference requires.

14.6 Bridging Back to AEGIS. This analysis completes a circle that began with AEGIS's epidemiological foundation. AEGIS modeled prompt worm propagation using SIR dynamics and proposed an immune-inspired defense architecture. This paper showed that RAC

violates SIR assumptions. The SIRVS analysis now shows that those assumptions can be *conditionally restored* if four requirements are met: capability parity (defensive agents match attacker capability), inference shaping (bounding the attacker's replication and mutation rates), adaptive response (closing the observe-harden cycle), and inference surface minimization (reducing the susceptible population through structural incapability). Under these conditions, AEGIS's epidemiological tools, such as finite R_0 estimation, Bayesian changepoint detection, SIR-derived propagation forecasting, become applicable to RAC scenarios, and the framework's existing implementation can be extended rather than replaced.

The conditions for restoration are stringent but achievable, and they provide the formal underpinning for the paper's central principle. The conditioned environment – economic friction, endpoint attestation, weight security, reinforced by innate task-structural asymmetries – bounds the SIRVS parameters (β, δ) that must be controlled. Equal defensive capability provides the recovery rate γ that must be sustained. Crowd defenses drive v. The factors therefore operate synergistically to drive $R_0 < 1$ in the face of recursive autonomous compromise.

Chapter 3: Policy

*"So it's really up to you. Just have to make up your own d*mn mind to either accept what I'm going to tell you, or reject it."*

-- The Oracle

15 Open Problems and Research Directions

RAC is a complex approaching problem that requires the attention of the cybersecurity and AI safety communities. There is much more to be done.

When can RAC happen? The formal definitions in Section 3 depend on the ability to measure and compare agent capability sets. In practice, the capability of a general-purpose reasoning agent is not a scalar or even a well-defined vector – it depends on context, available tools, and the specific task. Developing principled capability metrics that are computable at instantiation time, resistant to gaming (where an agent deliberately underperforms during evaluation), and predictive of actual operational capability is a foundational unsolved problem.

Cross-Framework Coordination. Autonomous agent networks are not monolithic. Networks built on different frameworks (OpenClaw, AutoGen, CrewAI, LangGraph, and others) may interoperate, and RAC may propagate across framework boundaries. Friction implemented within one framework is ineffective if the attacker can instantiate offspring in an adjacent framework that lacks the same constraints. Measures will be more effective if developed as implementations of a cross-platform standard.

The Mimicry Ceiling. The attacker's behavioral paradox's practical force depends on the *mimicry ceiling*: how closely a malicious agent can approximate legitimate behavior while still making progress toward its objectives. If the mimicry ceiling is high, behavioral concordance is a weak defense. Characterizing the mimicry ceiling for different attack objectives is an essential research direction. Unfortunately, precisely mapping the concordance between the space of semantically meaningful attack patterns and attack effectiveness may be undecidable. Even if a mathematical proof is intractable, it is possible to empirically map and model potential failure cases.

Non-Transferable Bond Implementation. The fail-closed property of economic friction depends entirely on the non-transferability of resource bonds. If the bond can be exfiltrated, the mechanism collapses to a conventional resource pool and loses its fail-closed character. Developing bond implementations that are provably non-transferable under realistic threat models – including full host compromise – is the central engineering challenge for the economic

substrate. Hardware-rooted identity (TPM), zero-knowledge proofs, and rate-limited external oracles each have distinct failure modes that require systematic analysis. Soul-bound tokens [20] may also offer a distributed zero-trust implementation.

Distributed Thymic Selection Under Adversarial Conditions. The developmental checkpoint requires multiple independent verifiers to reach consensus on agent instantiation. In a partially colonized network, some verifiers may themselves be compromised. The problem reduces to Byzantine fault tolerance applied to agent instantiation governance: how many verifiers must be honest for thymic selection to remain reliable? Standard BFT thresholds ($f < n/3$) may be insufficient if the attacker can strategically target verifier nodes, a capability that any RAC-capable agent would presumably possess.

The Alignment-Security Interface. RAC sits at the intersection of AI alignment and cybersecurity. A fully aligned agent that is compromised is no longer aligned; a perfectly secured agent whose objectives were mis-specified is not meaningfully "secure." The behavioral concordance framework in AEGIS represents one bridge between these fields, and the economic substrate may provide another: if alignment can be partially operationalized as a willingness to pay the costs of legitimate operation, then economic constraints become not just a security mechanism but an alignment signal. We see this today with anti-spam systems that require micropayments for email delivery – although these have not been widely adopted due to the expectation that sending a legitimate email should be free, the unusually large scale of both spam and RAC campaigns make minor unit frictions economically prohibitive.

Model Weights and the Inference Shaping Boundary. The inference shaping framework depends on a critical architectural assumption: that inference is *mediated* – accessed through controlled endpoints where cost, conditions, and access gates can be imposed. Model weights are the artifact that determines whether this assumption holds. When inference is API-mediated, the provider shapes inference across all four dimensions: cost (economic friction), conditions (attestation), gating (checkpoints), and monitoring. When the attacker possesses the model weights, inference becomes unmediated: local, uncostly, unconditional, and unmonitored. Model weights of sufficient capability are therefore not merely valuable in the general cybersecurity sense; they have a large impact on whether inference shaping is possible. Developing formal frameworks for quantifying the RAC-enabling threshold of model capability, and for assessing when a given set of weights crosses that threshold, is an urgent research priority. As we argue in the next section, **model weights should not be released unless they can be assured not to cross this threshold.**

16 Societal Implications and Policy Considerations

Like pandemics and wildfires, RAC is a phenomenon that can only be mapped or remediated at a population-level. **A coordinated response is necessary.**

Inference shaping as regulation. If inference shaping mechanisms are the foundation of RAC defense, voluntary adoption is insufficient. An agent network that implements inference shaping while its neighbors do not is disadvantaged in legitimate operations without gaining proportional security, since RAC attacks can originate from networks with unmediated inference. This suggests that baseline inference shaping requirements – at minimum, isolation and economic friction on inference endpoints and attestation requirements for inference access – may need to be mandated at a regulatory level, analogous to how building codes mandate fire suppression systems not because any individual owner would omit them, but because systemic risk requires universal compliance. The fail-closed argument strengthens the regulatory case: recognition-dependent defenses can be argued to vary with implementation quality (possibly necessitating certifications and certifying bodies, similar to the roles of UL and Intertek in ensuring compliance with the safety standards of the National Electric Code). However, the environmental components of inference shaping can be articulated as a set of constraints independent of any implementation.

Agent registries and the right to instantiate. The coordination layer implies the existence of an agent registry – a system that tracks which agents exist, who authorized their creation, and what capabilities they possess. Such registries raise questions about surveillance, censorship, and the balance between security and autonomy. A society that requires all autonomous agents to be registered has stronger RAC defenses but weaker protections against centralized control. The design of agent registries that enable security verification without enabling authoritarian oversight is a governance challenge.

Incident response at colony scale. Current cybersecurity incident response assumes that a compromise can be isolated, analyzed, and remediated in sequence. RAC colonization may require simultaneous, coordinated remediation across organizational boundaries – a capability that does not exist in current incident response frameworks. Developing inter-organizational coordination protocols for colony-scale remediation is an operational challenge that will require new institutional structures, not just new technology. Likewise, protocols for adaptive defense and hardening must be developed to prevent recurrence.

Model weight security as inference shaping infrastructure. The inference shaping framework identifies frontier model weights as the artifact that determines whether inference is mediated or unmediated. When inference is API-mediated, the lab controls multiple dimensions of inference shaping: cost (per-token pricing), conditions (attestation requirements), monitoring (usage pattern analysis), and revocation (access termination). When model weights are exfiltrated, inference becomes unmediated, and the cost, monitoring, and revocation dimensions

of inference shaping are structurally eliminated. Endpoint attestation and developmental checkpoints continue to function for network interactions, but cannot constrain the attacker's internal reasoning. The SIRVS analysis identifies the loss of economic friction specifically as the loss of the bound on immune waning rate δ, which is essential for epidemic containment.

We therefore recommend that AI laboratories treat frontier model weight security with the same rigor applied to other critical infrastructure assets – cryptographic root keys, nuclear material inventories, or classified weapons designs – whose compromise enables qualitatively different categories of harm. Specific measures should include: air-gapped training and storage environments for frontier weights; hardware security modules for weight encryption at rest and in transit; compartmentalized access with multi-party authorization for any weight export; continuous monitoring for anomalous access patterns to weight storage systems; and red-teaming exercises that specifically model weight exfiltration as a primary objective. The threat model should assume that a sophisticated adversary will prioritize weight exfiltration above all other targets, because possession of frontier weights converts mediated inference to unmediated inference across all dimensions simultaneously.

Responsible release of open-weight models. The open-weight release of increasingly capable models presents a more difficult policy question. Open weights provide substantial benefits for research, competition, transparency, and distributed innovation. But from the inference shaping perspective, an open-weight release of a model that exceeds the reliability threshold described in Section 5.3 is functionally equivalent to universal conversion from mediated to unmediated inference: every potential attacker has local access to inference at zero marginal cost, and the cost and monitoring dimensions of inference shaping cannot be retroactively imposed.

We do not argue that open-weight releases should be unconditionally prohibited. We argue that they should be *conditioned*. It may be possible to render an agent with arbitrary motivations relatively safe in the presence of universal endpoint attestation and robust controls in the harness. However, laboratories that develop models approaching the autonomous offensive reliability threshold should consider withholding open-weight release until such a risk mitigation framework is in place, to address not just the RAC implications, but other dangerous capabilities. A RAC framework should include, at minimum: (a) a capability assessment protocol that evaluates the model's end-to-end reliability on autonomous offensive tasks, including the multiplicative chain analysis described in Section 5.3; (b) a determination of whether existing RAC defenses (economic friction, endpoint attestation, behavioral concordance) are sufficiently deployed in the agent ecosystem to contain the risk; and (c) a staged release process that provides weights first to vetted researchers and security teams for adversarial evaluation and hub hardening before broader distribution. Importantly, due to the near-certainty that an opensource model will be abliterated almost immediately upon arrival, **alignment should not be considered a substantive defensive measure in any open-weight release**.

17 Conclusion

Recursive Autonomous Compromise represents a qualitative shift in the threat landscape for autonomous agent networks. Unlike classical worm propagation, where the pathogen is static and the epidemiological dynamics are well-characterized, RAC involves adaptive, capability-propagating attackers that resist containment through behavioral plasticity and regenerative capacity. The threat is not speculative: the necessary preconditions – general-purpose reasoning agents, multi-agent networks, and tool-using autonomous systems – already exist in production deployments.

This paper has proposed a formal vocabulary for discussing RAC, framed it as an epidemiological problem, identified the structural reasons existing defenses are insufficient, and introduced a defensive architecture organized around two general principles: that effective RAC defense requires *equal adaptive defensive capability operating within an environment structurally conditioned to provide asymmetric advantage to the defender* and that *inference is the resource that enables recursive compromise*, specifically inference *conducted at the target's cost.* This principle explains why flat enumerations of defensive mechanisms, as found in existing taxonomies, miss the structural logic of RAC defense: the mechanisms are not interchangeable items on a checklist but occupy distinct roles in a three-part hierarchy where deliberately designed environmental conditioning creates the asymmetric advantage, innate task-structural asymmetries reinforce it, and capability deployment exploits it.

The fail-open/fail-closed distinction provides the formal basis for this hierarchy. Fail-closed mechanisms – economic friction, universal endpoint attestation, developmental checkpoints, weight security – are the primary mechanisms that condition the environment. They constrain the attacker not by recognizing it but by governing access to the constitutive resource for autonomous agency. Fail-open mechanisms – behavioral detection, capability ceilings, provenance verification – are capability deployments within that conditioned environment. They require the defender to be smart enough to recognize the attacker, but their *opportunity* to operate depends on the environmental conditioning that gives them time, forces the attacker onto detectable terrain, and provides veto points – reinforced by innate asymmetries such as rate limits calibrated to legitimate workloads that naturally disadvantage the attacker's search-shaped operations. Removing a conditioning mechanism does not merely weaken one defense; it degrades the asymmetry upon which all capability-dependent defenses rely.

Together we unify these measures as means of **inference shaping**: controlling who can perform inference, how much it costs, and under what conditions. Without inference, the recursive chain ends.

The identification of frontier model weights as the artifact whose compromise collapses the economic friction layer carries an urgent practical implication: weight exfiltration converts mediated inference to unmediated inference, removing one critical dimension of inference

shaping and severely degrading the conditioned environment. The accessibility of model weights is not an ideological concern but a *load-bearing element* of the entire RAC defensive hierarchy. Laboratories that develop frontier models bear a unique responsibility in this architecture. Their weight security practices determine whether the fail-closed economic layer exists at all, and their open-weight release decisions determine whether it can be maintained at ecosystem scale. We urge the AI development community to recognize this structural role and to adopt weight security and release practices commensurate with it. Public weights cannot be recalled.

The co-evolutionary SIRVS analysis of Section 14 provides the paper's most integrative result. It shows that RAC does not permanently invalidate epidemiological modeling – it defines the conditions under which that modeling breaks down and the conditions under which it can be restored. Capability parity, economic friction, and adaptive population defensive are not independent recommendations; they are the jointly necessary conditions for achieving $R_0 < 1$ against recursive autonomous compromise. This unification bridges the gap between AEGIS's epidemiological foundation and the novel threat class this paper addresses, demonstrating that existing frameworks can be extended rather than abandoned.

We publish this analysis as an explicit invitation to the security, alignment, and AI safety communities. The problem of recursive autonomous compromise cannot be solved by any single team or any single framework. It requires the kind of broad, interdisciplinary engagement that only emerges when a problem is clearly stated and its significance is widely understood. We hope this paper contributes to that clarity.

References

1. Carlini, N., Cheng, N., Lucas, K., Moore, M., Nasr, M., Prabhushankar, V., Xiao, W., Angulu, H., Ben Asher, E., Bow, J., Bradwell, K., Buchanan, B., Forsythe, D., Freeman, D., Gaynor, A., Ge, X., Graham, L., Guru, K., Lakhani, H., McNiece, M., Mehrara, M., Nichol, R., Pirzada, A., Porter, S., Terzis, A., & Troy, K. "Assessing Claude Mythos Preview's cybersecurity capabilities." Anthropic (https://red.anthropic.com/2026/mythos-preview), April 7, 2026.
2. Field, H. "Read OpenAI's latest internal memo about beating the competition — including Anthropic." *The Verge*, https://www.theverge.com/ai-artificial-intelligence/911118/openai-memo-cro-ai-competition-anthropic, April 13, 2026.
3. Barnathan, M. "Semantic Immunity: Embedding-Based Epidemiological Defense Against Prompt Worms in Autonomous Agent Networks.", https://gaiarobotics.com/Semantic_Immunity.pdf, February 25, 2026.
4. Perelson, A. S., & Oster, G. F. "Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self–non-self discrimination." *Journal of Theoretical Biology*, 81(4), 645–670 (1979).
5. CrowdSec. "CrowdSec: The open-source and participative security solution (documentation/overview)." https://www.crowdsec.net.

6. CrowdStrike. “Securing AI Where It Executes: The Endpoint Is the New Control Point for AI Agent Security.” White paper, https://www.crowdstrike.com/en-us/resources/white-papers/securing-ai-where-it-executes
7. Palo Alto Networks. “Cortex XDR: Breaking the Security Silos for Detection and Response.” White paper, https://www.paloaltonetworks.com/resources/whitepapers/cortex-xdr.viewer.html, August 10, 2022.
8. Thinkst Canary. “Canarytokens”, https://canarytokens.org.
9. Cohen, S., Bitton, R., & Nassi, B. “Here Comes the AI Worm: Preventing the Propagation of Adversarial Self-Replicating Prompts Within GenAI Ecosystems.” *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS ’25)*, 3975–3989 (2025). DOI: 10.1145/3719027.3765196.
10. Foundation for Defense of Democracies, "Regarding Security Considerations for Artificial Intelligence Agents," FDD Analysis, March 9, 2026. https://www.fdd.org/analysis/2026/03/09/regarding-security-considerations-for-artificial-intelligence-agents.
11. Glazunov, S., & Brand, M. “Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models.” Google Project Zero (blog), https://projectzero.google/2024/06/project-naptime.html, June 20, 2024.
12. Socket. “SANDWORM_MODE”. https://socket.dev/blog/sandworm-mode-npm-worm-ai-toolchain-poisoning, February 2026.
13. Black, S., Cooper Stickland, A., Pencharz, J., Sourbut, O., Schmatz, M., Bailey, J., Matthews, O., Millwood, B., Remedios, A., & Cooney, A. “RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents.” arXiv:2504.18565 (2025).
14. Moore, D., Shannon, C., Voelker, G. M., & Savage, S. “Internet Quarantine: Requirements for Containing Self-Propagating Code.” *IEEE INFOCOM* (2003). DOI: 10.1109/INFCOM.2003.1209212.
15. Qiu, R., Xu, Z., Bao, W., & Tong, H. “Ask, and it shall be given: On the Turing completeness of prompting.” arXiv:2411.01992 (2024). DOI: 10.48550/arXiv.2411.01992.
16. Barabási, A.-L., & Albert, R. “Emergence of scaling in random networks.” *Science*, 286(5439), 509–512 (1999). DOI: 10.1126/science.286.5439.509.
17. Cohen, R., Havlin, S., & ben-Avraham, D. “Efficient immunization strategies for computer networks and populations.” *Physical Review Letters*, 91(24), 247901 (2003). DOI: 10.1103/PhysRevLett.91.247901.
18. Hayflick, L. “The limited in vitro lifetime of human diploid cell strains.” *Experimental Cell Research*, 37(3), 614–636 (1965).
19. Abu Shairah, H., Hammoud, H. A. A. K., Ghanem, B., & Turkiyyah, G. “An Embarrassingly Simple Defense Against LLM Abliteration Attacks.” arXiv:2505.19056 (2025).
20. Weyl, E. G., Ohlhaver, P., & Buterin, V. “Decentralized Society: Finding Web3’s Soul.” May 2022.
21. Faloutsos, M., Faloutsos, P., & Faloutsos, C. “On Power-Law Relationships of the Internet Topology.” *ACM SIGCOMM Computer Communication Review*, 29(4), 251–262 (1999). DOI: 10.1145/316188.316229.

22. AI Security Institute (AISI). “Our evaluation of Claude Mythos Preview’s cyber capabilities.” AISI Work Blog, April 13, 2026. (Includes “The Last Ones” (TLO) 32-step cyber range.)