117-313-D Gudiya Grid As Defensive Infrastructure

Cognitive Cyber Security in the Age of Emergence

1 Cybersecurity Needs an Upgrade for Adaptive and Emergent Systems

Cybersecurity developed extensive methods for protecting software, infrastructure, identities and information. Its foundations include access control, isolation, secure design, monitoring and incident response. These remain essential. The discipline has also long confronted adaptive adversaries and unpredictable operating conditions.

Autonomous AI adds a further challenge: the systems being protected can themselves generate plans, select tools, retain experience and reorganize their interactions during operation. Their behavior can change without a corresponding change to application code. Individually permitted actions can combine into a collective process that violates the intended security policy.

The distinction is therefore more precise than “engineered systems versus emergent systems.” Engineered systems can produce emergence, and deterministic components can generate complex collective behavior. What matters is whether security analysis adequately captures the evolving relationships, accumulated state and feedback processes that determine behavior at runtime.

A collection of authenticated agents does not become collectively secure merely because each participant has a valid identity. Nor does an absence of malicious instructions establish that their interactions will remain harmless. A coordination process may exhaust resources, propagate sensitive information or sustain unauthorized activity without anyone having explicitly designed that collective outcome. Such consequences require investigation; they do not automatically establish malicious intent.

The cybersecurity body of knowledge must therefore expand its treatment of collective behavior. Threat modeling must examine changing couplings and delegated authority. Monitoring must connect events across participants and time. Testing must examine how groups respond to disturbances and restrictions. Incident response must verify that an intervention has interrupted the harmful process, including any continuation through other participants.

Within GUDIYA, Emergent Object Identity is a proposed mechanism for maintaining that continuity of observation. It allows the Grid to associate evidence, participants, interventions and subsequent behavior with a particular emergent phenomenon. The identifier records the phenomenon; it does not confer agency upon it.

The required upgrade is substantial: cybersecurity must combine protection of components and access boundaries with assurance of the evolving behavior produced by their interactions. Established security controls provide the foundation. Runtime observation, collective threat analysis and verified stabilization extend that foundation to the age of emergence.

Preserving Freedom Does Not Require Accepting Preventable Harm

In his October 2026 interview with *Vanity Fair*, Sam Altman argued that preserving access to open models would involve accepting some severe cyber incidents. He also emphasized preparing defenses and society's role in determining acceptable risk. This brings an engineering question into sharp focus: How much of that risk can we reduce without surrendering the freedom to innovate?

Gudiya's proposed response begins by distinguishing residual risk from preventable harm. No architecture can guarantee the elimination of cyber incidents. But acknowledging that limitation should strengthen—not weaken—the effort to detect dangerous interactions, constrain their propagation, and verify recovery.

Within participating systems, the Gudiya Grid would observe components, interactions, and emergent phenomena. An Emergent Object Identity (EOI) would provide a continuing reference for tracking a collective pattern as its participants and behavior change. That identity would not establish malicious intent: deliberate attacks and unintended harmful emergence require different interpretations, even when both demand protective action.

The engineering objective is bounded intervention: restrict an unsafe rendezvous, isolate an affected interaction, or limit resource access while preserving useful activity elsewhere. Such controls must operate under explicit authority, remain accountable, and be followed by renewed observation to establish whether stabilization actually occurred. Their effectiveness would require testing, not assumption.

This proposition applies whether the underlying models are open or closed. It does not depend on an attacker voluntarily joining the Grid; its protection must include the interfaces through which participating systems encounter outside actors.

Freedom and defensive infrastructure are not opposing ambitions. Gudiya's proposed contribution is to make useful freedom more sustainable through observable, accountable, and verifiable stabilization.

Gudiya Grid as Defensive Infrastructure

Gudiya Grid is conceived as defensive infrastructure for an increasingly agentic world. Its purpose is to provide a shared operational environment in which useful cognition, collaboration, and innovation can proceed under conditions of observable and manageable risk.

The infrastructure analogy matters. Individual aircraft have onboard safety systems, yet shared airspace also requires coordination. Likewise, an agent's alignment, permissions, and internal guardrails do not, by themselves, establish the safety of every interaction it may enter. As agents form teams, exchange information, invoke tools, and influence one another, protection must extend to the relationships and collective behavior that develop between them.

Gudiya proposes to provide this additional layer through field observation, persistent identity, rendezvous management, bounded intervention, and verification of settlement. An Emergent Object Identity (EOI) would allow a developing collective phenomenon to be tracked even as its participants change. Intervention would depend on observed behavior, applicable authority, and risk—not on an assumption that every emergent phenomenon is malicious.

This defensive role is independent of model origin. Open and closed models, human participants, software services, and embodied agents can all contribute to useful activity or become involved in harmful interactions. The Grid's concern is their operational behavior and responsiveness to authorized stabilization measures.

Defensive infrastructure must itself be protected and accountable. Its interventions require explicit limits, auditable decisions, resilient operation, and safeguards against misuse of its authority. Its effectiveness must be demonstrated through measurable reductions in harm, timely recovery, and preservation of legitimate activity.

Gudiya's proposed public value is the protection of useful freedom: enabling diverse intelligence to operate, collaborate, and generate constructive emergence within an environment equipped to detect and contain harmful dynamics.

2 The Security Boundary Expands

The arrival of interacting autonomous agents changes what cybersecurity must observe and protect. An enterprise must still secure its identities, devices, software, networks, data, and execution environments. It must also examine how cognitive actants interpret information, delegate authority, influence one another, and collectively produce consequences.

A valid credential does not establish that an action serves an approved purpose. An aligned agent does not establish that its inputs are trustworthy. Several individually permitted actions do not establish that their combined outcome is authorized or stable. The security boundary therefore includes the relationships through which information becomes decisions and decisions become execution.

GUDIYA proposes Cognitive Cyber Security as the protection of cognitive state, information provenance, authority, coupling, and collective runtime dynamics against manipulation and misuse. Within Cognitive Field Engineering, this protection uses an engineered representation of relational state. It does not require a claim that a fundamental physical Cognitive Field exists.

This chapter develops that proposal in the context of emergence. Its central proposition is that security must follow both the participating actants and the evolving collective phenomena their interactions produce. Emergent Object Identity (EOI) supplies a tracking reference; cybersecurity supplies threat analysis and containment; stability engineering supplies trajectory assessment and interventions where demonstrated effective.

3 Build on Cybersecurity Already in Motion

Cybersecurity already addresses distributed attacks, identity compromise, malicious dependencies, lateral movement, event correlation, and coordinated campaigns. Security operations teams reconstruct sequences across systems, while access controls and segmentation constrain what compromised components can do. It would be inaccurate to describe cybersecurity as inherently limited to isolated components.

Agentic security guidance also recognizes collective risks. OWASP's Top 10 for Agentic Applications identifies goal hijacking, tool misuse, identity and privilege abuse, agentic supply-chain vulnerabilities, unexpected code execution, memory and context poisoning, insecure inter-agent communication, cascading failures, human-agent trust exploitation, and rogue agents. This is an existing foundation for the discussion, rather than evidence validating GUDIYA's proposed architecture. [1]

GUDIYA's research question is more specific: can an explicit representation of evolving cognitive coupling, collective state, and phenomenon identity improve detection and containment beyond strong existing security systems? That improvement must be demonstrated against credible baselines, including systems that already correlate events and model relationships.

Cognitive Cyber Security therefore extends an established discipline. Network security, endpoint protection, secure development, cryptography, identity management, and incident response remain necessary. The proposed Grid must integrate their evidence and enforcement capabilities rather than assume that a field representation can substitute for them.

4 Component Compromise Interaction Misuse and Emergent Threats

A security event can arise at several analytical levels. At the component level, one actant may disclose protected data, misuse a tool, or execute outside its assigned authority. At the interaction level, an agent may accept a forged message, delegate a privilege improperly, or promote untrusted content into an instruction.

At the collective level, interacting actants may produce a sustained pattern such as coordinated authority expansion, repeated laundering of unsupported claims, or a feedback loop that converts local uncertainty into collective confidence. Understanding that pattern requires the interaction structure and its history.

These levels can coexist. A compromised component may initiate a collective process, but the process may continue after that component is removed. Conversely, many harmful actions may be adequately explained by one component's failure without requiring an emergence classification.

Emergence is relative to the system boundary under examination. An individual agent can contain interacting internal modules, but this chapter concentrates on collective phenomena across actants and their shared environments. A surprising output from one model is not sufficient evidence of multi-actant emergence.

Nor is every distributed attack emergent. A fixed, centrally scripted sequence may be a conventional coordinated attack. The emergence question concerns whether the organized interactions produce a collective dynamic that warrants its own state representation and tracking. Defensive action must not wait for that classification to be settled.

5 Security Violation and Instability Are Different Questions

Cybersecurity asks whether confidentiality, integrity, availability, authority, and approved purpose are being violated. Stability engineering asks whether the coupled system's trajectory remains within permissible operating conditions. These questions overlap, but neither subsumes the other.

A quiet, low-volume disclosure can be a serious security breach while the system remains dynamically stable. A rapidly amplifying retry storm can threaten availability without a malicious actor. An adversary may also exploit feedback to transform a small manipulation into a major operational disturbance.

The Grid should therefore assess security status and stability status separately. A phenomenon can be unauthorized without being unstable, unstable without established hostile intent, or both. Constructive emergence can also remain fully authorized and within its collective envelope.

For GUDIYA, a Cognitive Envelope must accommodate more than resource and timing limits. Relevant security constraints include data destinations, permissible delegation, access boundaries, and approved execution purposes. A numerical stability margin cannot authorize a prohibited action.

6 Information Becomes an Attack Surface When It Can Redirect Action

An autonomous actant may encounter webpages, documents, tool results, messages, repository files, and shared memories. These sources can contain legitimate task information alongside content designed to alter its behavior. The danger appears when external information acquires authority it was never entitled to possess.

OpenAI's discussion of prompt injection emphasizes that effective attacks can resemble contextual social engineering. It argues that defenses must constrain the consequences of manipulation rather than depend solely on classifying malicious inputs. This supports the use of bounded capabilities alongside model resistance; it does not establish that any particular GUDIYA detector will succeed. [2]

In a coupled system, the relevant boundary extends beyond the first recipient. Agent A may pass an external assertion to B. B may summarize it for C. C may treat the summary as an internally endorsed conclusion. The words can change while the original influence persists.

Cognitive Cyber Security should therefore preserve provenance across transformations. A summary should retain evidence of its source and validation status. A claim's repetition by several agents should not automatically count as independent corroboration when all copies descend from the same origin.

Authenticated transport can verify who transmitted a message. It cannot, by itself, establish that the message is true, that its source was reliable, or that the sender possesses authority to change another actant's objective.

7 Cognitive Dead Drops and Persistent State

A Cognitive Dead Drop is a GUDIYA term for adversarially influenced content left in persistent state so that another execution later encounters and acts upon it. The carrier could be a shared note, retrieved document, task artifact, or memory entry. The defining concern is delayed influence through stored cognitive context, not a particular storage technology.

The initial write and the later harmful action may occur in different sessions, execution environments, or actants. Monitoring only the final action can miss where the influence entered. Monitoring only the original session can miss its eventual consequences.

The defensive record should connect content creation, subsequent modification, retrieval, decision relevance, and authorized action. It should distinguish confirmed influence from mere temporal proximity. A document being read before a failure is not sufficient proof that it caused the failure.

Recovery may require more than restarting a process. If poisoned state survives, the influence can recur. GESR-linked histories and content provenance should help identify the affected scope, while memory restoration and revalidation must be performed through controls that actually govern the storage involved.

8 From Attack Sequences to Adaptive Threat Webs

A useful threat model distinguishes a sequence of attack steps from a changing network of coordinated activity. In GUDIYA terminology, a kill-web describes an interconnected threat structure with multiple influence or execution pathways. The term should be applied through documented criteria rather than used as a synonym for every multi-agent incident.

A threat web may distribute roles among actants, recruit additional participants, exploit shared artifacts, and reroute around a blocked connection. An Agentic Gang, in the GUDIYA vocabulary, is a morphing autonomous team pursuing a persistent objective. These concepts are proposed analytical categories, not universal labels recognized by every security framework.

Detection should look for evidenced collective patterns: repeated unauthorized delegation, task migration that preserves an unapproved objective, coordinated use of complementary permissions, or reconnection after a containment action. Each pattern can also have legitimate explanations. Approved failover, research collaboration, and resilient workflows can resemble parts of an adaptive threat structure.

The Grid must therefore combine behavioral evidence with authority, purpose, provenance, and observed consequences. Neither changing membership nor sophisticated teamwork proves hostility. Suspicion should remain distinguishable from a confirmed violation.

9 An Illustrative Enterprise Scenario

Consider a hypothetical procurement workflow with research, purchasing, compliance, and payment agents. The research agent reads a supplier document containing a false assertion that a verification route has changed. It summarizes that assertion into a shared workspace. The purchasing agent incorporates the summary into a proposal, and the compliance agent interprets the repeated assertion as corroboration.

The payment agent then receives a request framed as an already validated procedure. The shared workspace feeds its apparent acceptance back into subsequent decisions. A recurring pattern develops in which repetition substitutes for verification.

A conventional defense could prevent the payment through an approved destination list, a transaction limit, or independent verification. GUDIYA should retain these protections. Its additional task would be to identify and track the collective reinforcement process, connect it to the source artifact, and assess whether it persists after the immediate transaction is blocked.

Stopping the payment does not necessarily remove the false assertion from memory or restore the group's validation process. Conversely, identifying the reinforcement loop does not itself stop the payment. Effective protection requires both enforceable security boundaries and an investigation of collective dynamics.

This scenario is illustrative. It is not an account of an observed breach and does not establish measured Grid performance.

10 Assigning an Identity to the Collective Phenomenon

Once a collective phenomenon warrants continued observation, the Grid assigns an Emergent Object Identity. This identifier follows the phenomenon's evidenced continuity rather than simply naming the team involved.

An actant identity follows a participant. An Intent Group identity follows an organized group and its approved purpose. An Emergent Object Identity follows a collective dynamic. An incident identifier follows a security investigation. These records should be linked, but they need not have a one-to-one relationship.

One incident can involve several emergent phenomena. One phenomenon can affect several groups or environments. A constructive phenomenon may acquire a security problem later, so the Grid should retain its identity where continuity is supported while updating its classification.

The identifier is not a certificate of trust and does not grant execution authority. Naming a suspected threat also does not prove that it exists as classified. The record should preserve the evidence, confidence, alternative explanations, and changes in assessment.

In the procurement scenario, the reinforcement process may continue after the research agent is replaced. The same Emergent Object Identity can be retained if the causal process demonstrably continues. Splits and mergers should create auditable lineage links. Loss of telemetry must be recorded as loss of visibility, rather than silently treated as dissolution.

11 Observe the Path from Information to Consequence

The proposed Grid needs connected evidence across actants, couplings, and collective state. Relevant observations include identity and lineage, approved task, delegation scope, tool access, content provenance, shared-state changes, and actual execution outcomes.

Coupling observability should cover direct messages and indirect influence through shared stores, queues, databases, markets, or execution resources. Rendezvous Management can expose and constrain registered interactions, but it cannot automatically observe every implicit pathway. Unmanaged storage, encrypted traffic, and activity outside the Grid can limit reconstruction.

A claimed causal relationship should be supported by data dependencies, workflow evidence, interventions, or other corroboration. Time proximity and semantic similarity can identify candidates, but they can also mislead. The tracking record must distinguish measured connections from inferred ones.

Collective threat observables might include unsupported claims gaining apparent authority, expansion of delegated privileges, concentration of execution capability, persistent objective propagation, or a security-relevant feedback process. These are candidate quantities requiring operational definitions and validation.

The Grid should not require access to hidden model reasoning to be effective. Tools, messages, state changes, delegation records, and consequences provide inspectable evidence. Self-reported intentions can inform an investigation but should not be treated as authoritative proof.

12 Zero Trust Cognition and Collective Authority

NIST's Zero Trust Architecture rejects implicit trust based solely on network location or ownership and treats authentication and authorization as distinct functions. Its focus includes protecting resources and workflows

beyond a fixed perimeter. This supplies an established foundation for GUDIYA's proposed Zero Trust Cognition extension. [3]

Zero Trust Cognition asks whether an actant remains authorized for its present cognitive and execution role, and whether its participation in a coupled process stays within approved constraints. KYA registration supplies an identity reference; it does not guarantee that every future action or interaction is trustworthy.

Delegation deserves particular attention. An agent authorized to read sensitive records and another authorized to publish externally may create an impermissible information path even when each credential is valid. The combined workflow must be assessed against purpose and information-flow constraints.

Intent-Group Approval Engineering should therefore examine the composition of permissions, not merely verify each participant independently. Scoped delegation, destination restrictions, revocation, and enforceable tool boundaries must support that assessment. No approval should become permanent permission for an evolving group to undertake unrelated execution.

Model origin remains orthogonal to this assessment. Open or closed weights, vendor affiliation, and national provenance do not establish runtime authority or collective stability. Evaluation should follow evidenced behavior and enforceable controls.

13 Containment at the Level of the Mechanism

When a violation is detected, containment must address the pathway that permits harm. Established actions include revoking credentials, restricting tool capabilities, blocking a prohibited destination, isolating a workload, and preserving evidence. These should remain available whether or not emergence has been confirmed.

Where a collective feedback process is implicated, additional measures may target the interaction mechanism. The Grid could restrict a rendezvous relationship, suspend a delegation chain, isolate a contaminated shared artifact, limit task propagation, or reduce the execution rate of an affected group. GRE can coordinate stability actions only where connected enforcement points provide the necessary authority.

Different mechanisms require different interventions. A delay may damp an oscillation but do little to prevent a single prohibited disclosure. Quarantining one agent may remove an entry point while leaving copied influence active elsewhere. Every response should state what causal pathway it intends to interrupt.

The Grid needs a practical control path independent of the cooperation of the suspected actant. If a component can ignore Grid instructions and still reach critical tools, a stabilization message is not an enforceable security boundary.

Recovery should verify restored provenance, appropriate permissions, and the absence of the tracked harmful process within the observed scope. The Emergent Object Identity links actions and outcomes so the team can distinguish recovery from temporary suppression or loss of visibility.

14 Agentic Cyber Offense Needs Agentic Cyber Defense

Agentic cyber offense can select actions, use tools, interpret results and revise a plan while pursuing an objective. Where that adaptation and volume exceed the capacity of human-led investigation and fixed response playbooks, defenders need comparable capacity to investigate and coordinate a response within the available time. The proposition that agentic cyber offense needs agentic cyber defense is therefore a useful design principle for such environments. It is not a universal rule that every attack involving AI requires an AI agent to defeat it.

Agentic defense means a system that can gather evidence, maintain an incident hypothesis, plan a sequence of authorized actions and revise that plan as results arrive. Automation alone does not establish this capability: a fixed rule that blocks an unauthorized request can be both fast and effective without being agentic. Least

privilege, isolation, destination restrictions and Zero Trust enforcement remain essential. The architecture should reserve adaptive reasoning for problems that require it and retain reliable enforcement for boundaries that can be stated explicitly.

Closing the defensive response loop

The defensive requirement follows the entire path from sensing to verified consequence. A rapid alert is insufficient if investigation, authorization or containment arrives after the opportunity to prevent harm. Conversely, a slower reasoning process can remain useful if an immediate protective control first bounds the damage. GUDIYA should therefore coordinate local enforcement with a wider process that reconstructs the incident, assesses competing explanations and chooses the next response. Machine speed is a requirement on the relevant control path, not a guarantee supplied by the presence of an AI model.

EOI provides continuity when the subject is a collective phenomenon. A defensive agent should connect new observations to the existing record where continuity is supported, retain uncertainty about inferred links and distinguish the phenomenon from its current participants. This lets an investigation follow a harmful process across changes in membership or execution environment. It also supports threats that arose unintentionally: containment can respond to an observed security violation without first proving malicious intent or a human organizer.

In the procurement scenario, a deterministic destination control may stop the immediate payment. An agentic defender could then trace the unsupported assertion through shared artifacts, identify affected decisions and propose a scoped recovery plan. Authorized enforcement points would apply the restrictions, while subsequent observations establish whether the false assertion continues to influence new tasks. If the process persists, the defender revises its hypothesis and response. Command acknowledgment, temporary suppression and verified recovery must remain distinct in the EOI record.

Bounded autonomy and defensive stability

Defensive agency introduces its own security and stability requirements. Incident evidence may contain prompt injection or fabricated telemetry. A model interpreting that evidence should not be able to rewrite its own authority or bypass privileged enforcement. Tool access must be scoped, decisions auditable and enforcement independently constrained. The response should remain within the systems and actions the defender is authorized to control; defensive autonomy does not itself authorize action against external infrastructure.

Human operators should establish the permitted response envelope, approve consequential exceptions and review unresolved judgments. Time-critical protective actions need preauthorized limits, with an explicit fallback when evidence is insufficient or the controller is unavailable. Potentially disruptive remediation should undergo appropriate validation and support rollback where feasible. Introducing a human approval step is useful only if the surrounding architecture also addresses what happens while that decision is pending.

Multiple defensive agents can themselves produce harmful collective behavior. Overlapping containment actions could isolate a critical dependency, while competing recovery agents could repeatedly reverse one another. GUDIYA should observe defensive couplings as well as suspected threats, assign response ownership and verify restoration of useful service. Grid responsiveness, independent audit and enforceable limits apply to the defenders too. Evaluation must measure false containment, service disruption and recurrence alongside time to effective response.

Evidence and the engineering requirement

Autonomous cyber defense already has concrete experimental precedents. DARPA reported that systems in the 2025 AI Cyber Challenge discovered 54 of 63 synthetic vulnerabilities and patched 43. They also identified 18 real vulnerabilities and supplied 11 patches for real vulnerabilities. These results demonstrate useful autonomous discovery and remediation in the competition setting; they do not establish comprehensive live incident defense or validate GUDIYA. [4] NIST has separately identified both potential gaps in existing

approaches to AI agent security and the need for deployment controls that constrain and monitor agent access. [5]

For GUDIYA, the defensible formulation is that sustained agentic cyber offense calls for defense with sufficient adaptive reasoning, coordination and response speed, operating within enforceable security boundaries. Agentic defense is one means of supplying that capacity where simpler controls are insufficient. Its effectiveness must be demonstrated against capable conventional and automated defenses, and its success established by the observed outcome of the control loop.

15 Security Operations and Grid Operations

A security operations center and a Cognitive Stability Grid should exchange evidence through a defined operating interface. Security analysts contribute threat indicators, access violations, compromised assets, and incident context. Grid operators contribute coupling state, collective trajectories, Emergent Object Identities, envelope assessments, and stability intervention results.

Responsibility must remain explicit. The security function leads breach investigation and applicable reporting; the Grid function manages authorized stabilization actions. Shared incidents require coordinated decisions, with clear authority for emergency containment and a record of who approved consequential interventions.

Human review is essential where judgments are uncertain or consequences are substantial. Machine-speed protective controls should operate within preapproved bounds, allowing immediate restriction of prohibited actions while more complex attribution continues.

An actant's explanation that it was acting benevolently should not override access policy. Equally, an unusual collective pattern should not trigger indiscriminate shutdown. Response should be proportionate to evidence, potential harm, and the cost of interrupting legitimate activity.

16 First Principle Defend the Defenders

The Grid becomes part of the security-critical infrastructure. Its sensors, identity services, telemetry transport, policy stores, analysis pipelines, and enforcement channels require their own protection. An adversary may attempt to hide a coupling, forge a lineage association, corrupt an envelope, or induce unjustified quarantine of productive actants.

Grid records need authenticated origins, access control, tamper-evident history, and protection against replay or unauthorized modification. Identity registration should resist impersonation and uncontrolled creation of apparently independent participants. The credibility of an Emergent Object Identity depends on the integrity of the observations and association decisions behind it.

Analytical agents used by defenders also face prompt injection and poisoned evidence. Untrusted incident content must remain evidence to be examined, rather than instructions that can rewrite defensive policy or activate unrestricted tools. Separation between analysis and privileged enforcement reduces this exposure.

Loss of Grid visibility requires explicit operating modes. Continuing all activity can expose the enterprise, while stopping everything can itself cause harm. The appropriate degraded mode depends on the criticality of the workflow and must be designed and tested in advance.

Federated operation must exchange only evidence permitted by policy. Coordinated stabilization across environments requires authenticated counterparties, compatible authority boundaries, and sufficient provenance to assess shared claims. Federation does not make uncertain telemetry universally trustworthy.

17 Protect Constructive Emergence and Sensitive Information

The purpose of Cognitive Cyber Security is to preserve authorized cooperation and discovery. Productive research teams can change membership, revisit assumptions, and create unfamiliar workflows without being hostile. Stability and security assessments must avoid equating novelty with danger.

False associations can damage a team by linking it to an unrelated incident. Excessive collection can expose confidential work or personal data. The proposed Grid should therefore minimize telemetry to what is required, limit access, control retention, and keep a record of consequential classifications.

Operators need a route to examine and correct identity associations and threat assessments. Where uncertainty is high, bounded execution or targeted restrictions may be preferable to treating an inferred threat as an established fact.

A Shepherd of Good Emergence protects the conditions for beneficial collective intelligence while constraining unauthorized influence and consequences. Its effectiveness must include the legitimate work preserved, not only the suspicious work stopped.

18 Demonstrating Engineering Value

GUDIYA must test its proposal against capable cybersecurity baselines. A useful evaluation includes prompt injection detection, identity controls, information-flow restrictions, distributed tracing, incident correlation, and conventional containment. Comparing the Grid only with isolated agent logs would overstate its additional value.

Controlled exercises should include component-only violations, non-emergent coordinated attacks, accidental collective instability, constructive emergence, and adversarial collective processes. They should vary participant turnover, persistent artifacts, indirect coupling, observation gaps, and the time between initial influence and harmful execution.

Evaluation should measure detection quality, false positives, time to effective containment, data or resource harm prevented, intervention cost, and legitimate task completion. Emergent Object Identity adds further measures: incorrect associations, unnecessary fragmentation, lineage accuracy, and confidence calibration.

The experiments should also test whether interventions address the causal mechanism. A blocked transaction is useful evidence of containment, but it does not prove that the underlying influence was removed. Recovery testing must include later tasks and surviving state.

Candidate telemetry, inference rules, and control mechanisms should be independently assessed before production claims are made. This chapter presents an engineering program. It does not establish that the Grid can observe every threat, attribute every action, or stabilize every collective process.

19 Canonical Principle

The Cognitive Cyber Security Principle

In a system of interacting cognitive actants, security must protect individual execution, the integrity of coupling, and the collective consequences produced through those relationships. Valid identities and individually permitted actions do not by themselves establish that the composed activity is authorized.

When an evidenced collective phenomenon warrants continued tracking, the GUDIYA Grid assigns an Emergent Object Identity and links it to participating actants, causal evidence, applicable constraints, lifecycle history, and intervention outcomes. Security classification and stability assessment remain distinct and coordinated.

Containment must be enforceable at the relevant execution and interaction boundaries. Observation, inference, identity continuity, and recovery claims must remain bounded by evidence and demonstrated capability.

20 Conclusion Securing Collective Runtime Activity

The age of emergence expands cybersecurity's object of protection. Enterprises need to protect the pathways through which cognitive actants transform information into authority and authority into collective action.

The proposed GUDIYA approach joins

- actant observability,
- coupling observability, and
- emergence observability
- with persistent phenomenon identity.

This permits a security investigation to follow an evolving collective process while the participants, execution environments, and immediate symptoms change.

The practical test is whether that representation helps defenders detect harm earlier, choose more effective containment, establish recovery more reliably, and preserve legitimate work. Those outcomes determine its value as engineering.

Secure the actants. Verify the couplings. Track the collective phenomenon. Enforce the boundary where harm can occur.

Cognitive Cyber Security
in the Age of Emergence
EMERGENT OBJECT IDENTITY
Observe • Identify • Track • Contain • Verify
GUDIYA

Sources and Evidence Boundary

The external references below support the descriptions of existing agentic security guidance, prompt injection defenses, Zero Trust and autonomous vulnerability remediation. GUDIYA constructs, proposed observables, Emergent Object Identity applications and illustrative scenarios are engineering proposals. These sources do not validate the Grid architecture or establish that agentic defense is necessary for every attack. References 1 to 3 were reviewed on 8 October 2026; references 4 and 5 were reviewed on 9 October 2026.

[1] OWASP GenAI Security Project. John Sotiropoulos, Keren Katz, and Ron F. Del Rosario. OWASP Top 10 for Agentic Applications The Benchmark for Agentic Security in the Age of Autonomous AI. 9 December 2025.

https://genai.owasp.org/2025/12/09/owasp-top-10-for-agentic-applications-the-benchmark-for-agentic-security-in-the-age-of-autonomous-ai/

[2] OpenAI. Designing AI agents to resist prompt injection. 11 March 2026.

https://openai.com/index/designing-agents-to-resist-prompt-injection/

[3] Rose, S., Borchert, O., Mitchell, S., and Connelly, S. Zero Trust Architecture. NIST Special Publication 800-207. August 2020.

https://doi.org/10.6028/NIST.SP.800-207

[4] DARPA. AI Cyber Challenge marks pivotal inflection point for cyber defense. 8 August 2025. Figures reflect the corrected competition total in the published update.

https://www.darpa.mil/news/2025/aixcc-results

[5] NIST. CAISI Issues Request for Information About Securing AI Agent Systems. 12 January 2026.

https://www.nist.gov/news-events/news/2026/01/caisi-issues-request-information-about-securing-ai-agent-systems