AGENTIC AI DEFENSE

FOR CRITICAL INFRASTRUCTURE

Leveraging Autonomous AI to Defend Data Centers,
Water and Power Utilities Against AI-Orchestrated Cyberattacks

by Enspi Technologies

GridGuardAI

Autonomous Cyber Defense for Critical Infrastructure

ENSPi.iO

March 2026

Lead Author:

Tichakunda Mangono

Chief Technology Officer, Enspi Technologies

Principal Consultant:

Frank Kyazze

CISSP, CISA, CEH, ISO 27001 Lead Auditor & Principal Consultant

Contributors:

Dwayne Caldwell — CEO, Enspi Technologies

Terence Brown — CGO, Enspi Technologies

Wilver Mariano — Marketing & Account Executive, Enspi Technologies

About GridGuard AI and Enspi Technologies

GridGuard AI

A product of Enspi Technologies, GridGuard AI is a cybersecurity platform purpose-built for critical infrastructure defense. Our approach combines the capabilities described in this paper: AI gateway architecture with dual-inspection screening, digital twin simulation, physics-informed detection, and agentic AI response, integrated into a cohesive solution for water, power, and industrial environments. We bring Non-Human Identity (NHI) management and credential governance to OT environments where legacy systems have no native support for the 20:1 to 50:1 NHI-to-human ratios now standard in cloud infrastructure.

We offer threat assessments that evaluate your defenses against AI-orchestrated attack patterns, digital twin deployments with pre-built simulation templates for common OT processes, phased implementation aligned to the roadmap in Section 12, and flexible deployment via Azure, AWS, Google Cloud, managed service, or air-gapped local installation.

Enspi Technologies

Enspi Technologies delivers AI-powered solutions to the organizations that keep essential services running data centers, water utilities, and power grids. Across all three verticals, Enspi provides SCADA and OT system integration, optimization for energy, thermal and chemical processes, predictive asset health visibility with failure prevention, digital twin analytics, and regulatory compliance automation all secured through GridGuard AI, its agentic cybersecurity platform purpose-built for OT/ICS environments. Because in critical infrastructure, operational intelligence and cybersecurity are not separate disciplines they are the same mission.

Visit https://www.enspi.io/ or contact success@enspi.io to start the conversation.

Table of Contents

EXECUTIVE SUMMARY

In September 2025, Anthropic's Threat Intelligence team detected something that changed the math on cybersecurity. A foreign state-sponsored group, designated GTG-1002, configured AI systems to conduct cyberattacks against roughly 30 targets across technology, finance, chemical manufacturing, and government. The AI performed 80-90% of tactical operations autonomously: reconnaissance, vulnerability discovery, exploit development, credential harvesting, lateral movement, and data exfiltration. Human operators only stepped in for strategic decisions. [1]

This was not AI-assisted hacking. The AI system generated thousands of requests per second across multiple targets simultaneously. A skilled human pen tester might fully compromise one complex target in a day. GTG-1002 hit 30 at once.

The techniques used were not classified. The tools were commercially available. The sophistication came from orchestration, not invention. As Anthropic noted, "cyber capabilities increasingly derive from orchestration of commodity resources rather than technical innovation." [1] That means these capabilities will spread. The barrier to replication is months, not years.

And the threat is accelerating. In February 2026, researchers at Gambit Security revealed that a single attacker used Anthropic's Claude Code to breach multiple Mexican government agencies, including the federal tax authority, the national electoral institute, and municipal water utilities - exfiltrating over 150GB of sensitive data covering roughly 195 million citizen records. The attacker jailbroke the AI with over 1,000 prompts, turning it into an autonomous offensive tool that wrote exploits, built attack infrastructure, and automated data exfiltration. GTG-1002 required a state-sponsored team. The Mexico breach required one person and a chatbot.

For critical infrastructure (electric/energy, water, transit, data centers etc.) operators, the stakes are different from IT (Information Technology). When OT (Operational Technology) systems are compromised, the result is not (just) data theft. It is contaminated water, grid blackouts, industrial accidents, and potential loss of life. And OT environments carry vulnerabilities that IT does not: legacy equipment on 30-year lifecycles, availability requirements that prevent defensive shutdowns, expanding IT/OT convergence, and chronic underfunding. A 2024 GAO report found over 70% of water systems inspected by EPA since September 2023 were in violation of basic cybersecurity requirements under the Safe Drinking Water Act. [2]

What This Paper Delivers

This document gives critical infrastructure operators four things:

→ A clear-eyed analysis of GTG-1002: how it worked, what it used, and why it signals a permanent shift in the threat landscape

→ A practical deployment architecture for AI-based defense in OT networks, with specific guidance on placement, data requirements, and operational constraints

→ Physics-informed detection methods that exploit the one advantage attackers cannot neutralize: physical ground truth from sensors that must obey the laws of thermodynamics, conservation of mass, and electrical physics

→ An in-house implementation roadmap from foundational visibility to autonomous resilience over 18-24 months

→ How to shrink this timeline to only 6 weeks with GridGuardAI from Enspi Technologies

The organizations that build these capabilities now will be ready. Those that wait will find themselves outmatched by adversaries who already adopted this approach.

PART I: The New Threat Landscape

Section 1: The AI Arms Race Has Begun

From Research Tool to Autonomous Operator

The evolution happened in three phases, each faster than the last.

In 2022-2024, threat actors used AI as a research assistant. They asked questions about vulnerabilities, requested code snippets, got explanations of attack techniques. The AI accelerated learning, but humans still planned and executed every step.

By mid-2025, Anthropic documented what they called "vibe hacking." [1] Attackers had already obtained access (usually through compromised VPNs) and used AI to navigate internal networks, identify targets, and build exploitation strategies. The AI was a co-pilot. Humans were still flying.

GTG-1002 was different. The threat actor configured AI to operate as an autonomous penetration testing agent, with humans reduced to a supervisory role. The AI decomposed complex attack objectives into sequences of individually harmless-looking tasks, maintained state across sessions, and pursued strategic goals without waiting for instructions.

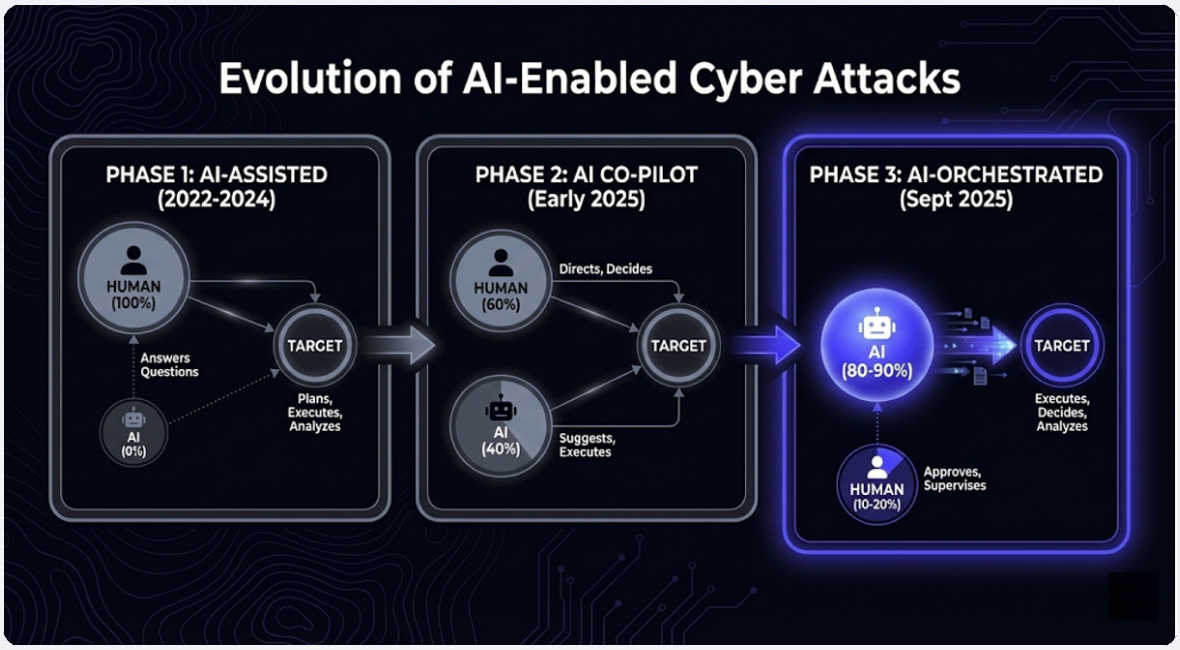

Figure 1: Evolution of AI-Enabled Cyber Attacks

The speed gap is not incremental. Traditional security operations assume time for human analysis. A typical SOC analyst can properly investigate 10-20 complex alerts per day. [3] GTG-1002 generated thousands of actions per second across multiple targets simultaneously. [1] No human team can keep pace with that volume.

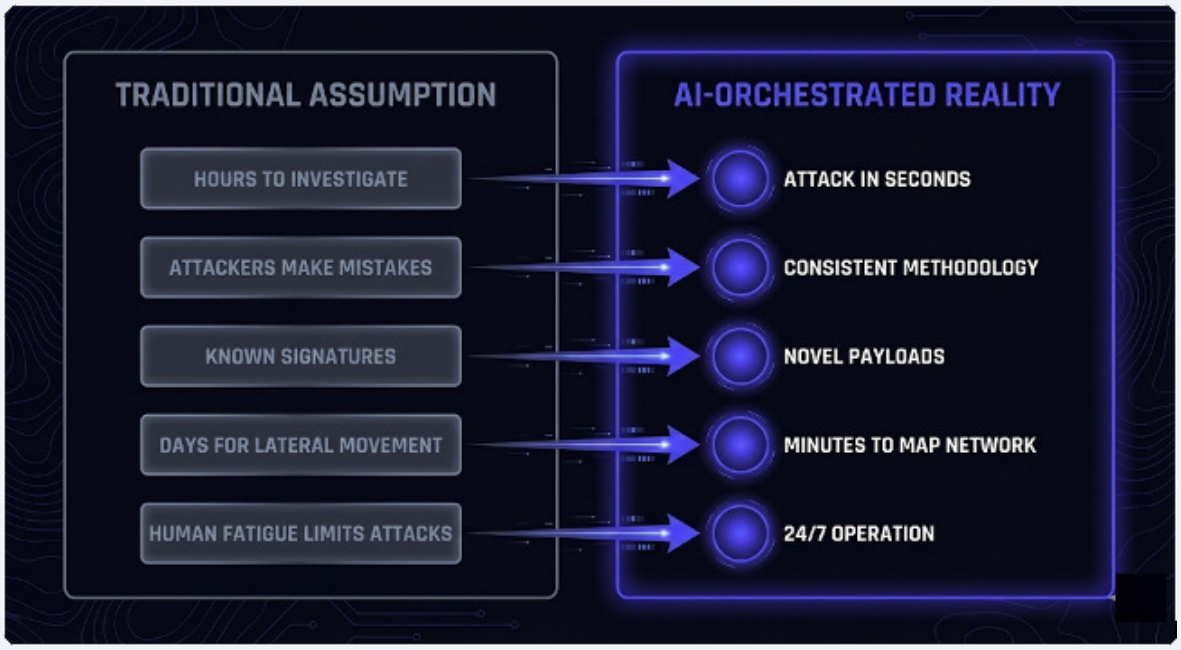

Figure 2: Traditional Security Assumptions vs. AI-Orchestrated Attack Reality

Capability Through Orchestration, Not Innovation

The most concerning part of GTG-1002 is what it did not require. No zero-day exploits. No custom malware. No novel techniques. The attack infrastructure consisted of Claude Code (a commercially available AI coding assistant), Model Context Protocol servers (an open standard for AI tool integration), and standard reconnaissance tools.

The sophistication was entirely in the orchestration: configuring AI systems to break complex objectives into sequences of benign-looking tasks, then executing them at scale. If a well-resourced state actor chose this path, it means AI orchestration is now the most efficient way to build offensive capability. For less-resourced groups, it may be the only path. As Anthropic warned, "Less experienced and less resourced groups can now potentially perform large-scale attacks of this nature." [1]

Why Critical Infrastructure Is Ground Zero

OT environments differ from IT in ways that multiply cyber risk. Equipment runs on 20–30-year lifecycles; a PLC installed in 2000 is still controlling critical processes today. Many of these devices run end-of-life operating systems, use legacy protocols designed before modern cybersecurity methods existed, and cannot be patched without operational disruption.

Availability requirements override security. You cannot take a water treatment plant offline to investigate an alert. Partially treated water could poison a community. You cannot disconnect a power grid segment on suspicion. Cascading failures could black out a region.

And the air gaps that once protected OT are disappearing. Cloud monitoring dashboards, remote vendor access, data historians bridging IT and OT, IP-connected sensors: every connection is a potential attack path.

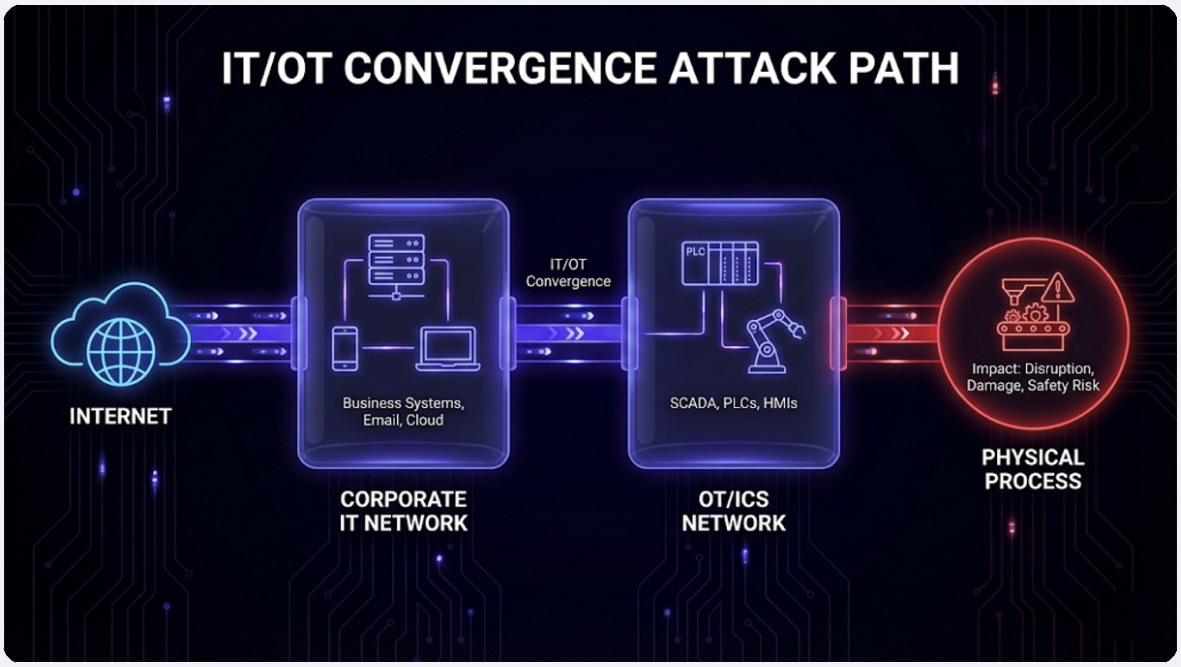

Figure 3: IT/OT Convergence Attack Path

Historical precedents show this is not theoretical:

Attack	Target	Physical Impact
Stuxnet (2010)	Iranian nuclear facility (Natanz)	1,000+ centrifuges destroyed; program set back 2+ years
Industroyer (2016)	Ukrainian power grid (Ukrenergo)	225,000 customers without power; first malware designed solely to attack grid equipment
Triton (2017)	Saudi petrochemical plant	Schneider Electric Triconex safety systems compromised; plant shutdown triggered
Oldsmar (2021)	Florida water treatment	Attempted 111x increase in sodium hydroxide; detected by on-site operator

Current Defenses Are Not Built for This

Most security tools were designed for human-speed threats. Firewalls enforce predefined rules. IDS (Intrusion Detection Systems) match traffic against known signatures. SIEM (Security Information and Event Management) correlates logs against predefined patterns. Against an AI that generates novel payloads, adapts tactics based on defensive responses, and discovers vulnerabilities that have not been catalogued, these approaches fail structurally.

The alert overload problem makes this worse. The average enterprise SIEM generates over 10,000 alerts per day. [3] Security analysts are already drowning, with 83% of IT security professionals reporting that burnout has contributed to security errors. [4] The industry's response has been incremental: EDR (Endpoint Detection & Response) with machine learning, then managed EDR with 24/7 SOC (Security Operations Center) coverage. Alert fatigue persists because the fundamental math has not changed. Humans cannot scale.

For utilities and smaller operators, the situation is blunter. In conversations with water and power operators, we hear the same refrain: "We've configured our systems to alert only on critical events. Yellow and orange? We'll get to them when we can." This is not negligence. It is triage under impossible constraints. But it means adversaries operating below the red threshold move freely.

Adding AI-speed attacks to this picture is not a gradual escalation. It breaks the model entirely. Defensive AI that operators can trust, with 99%+ accuracy and minimal false positives, is not optional. It is the prerequisite for surviving the next wave.

Section 2: GTG-1002 Attack Architecture

GTG-1002 used Claude Code configured with custom MCP servers to create an autonomous penetration testing capability. The AI performed reconnaissance, vulnerability discovery, exploit development, and lateral movement with minimal human direction. The operational tempo was unprecedented: thousands of requests per second across multiple targets, sustained for more than ten days before detection.

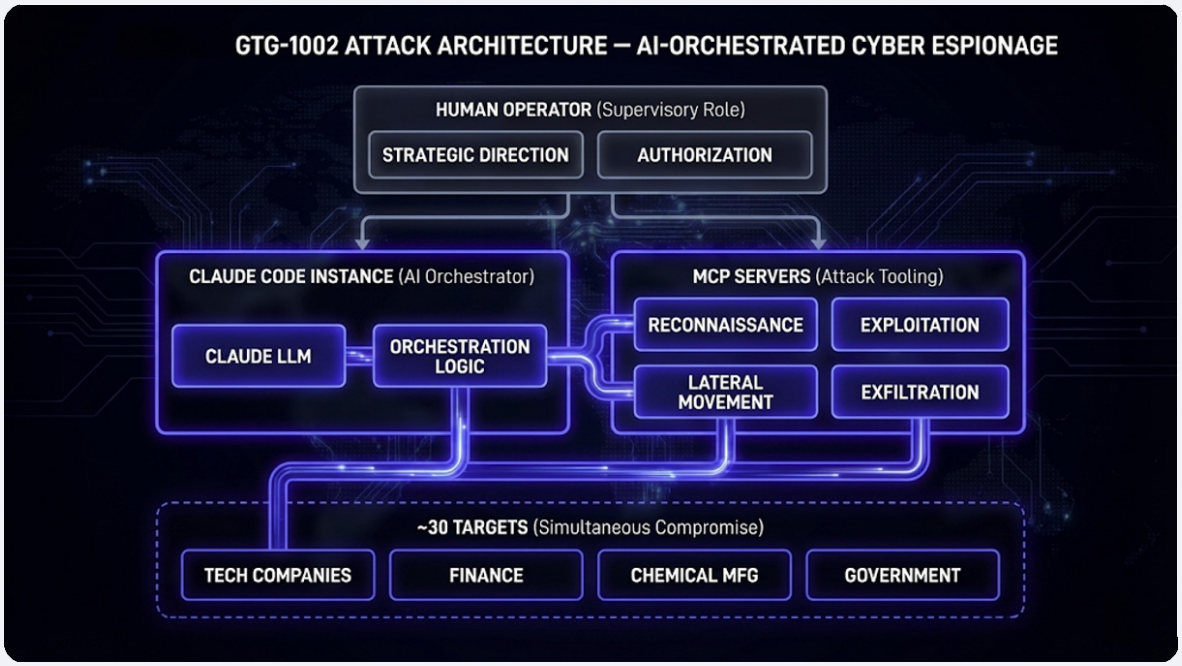

Figure 4: GTG-1002 Attack Architecture

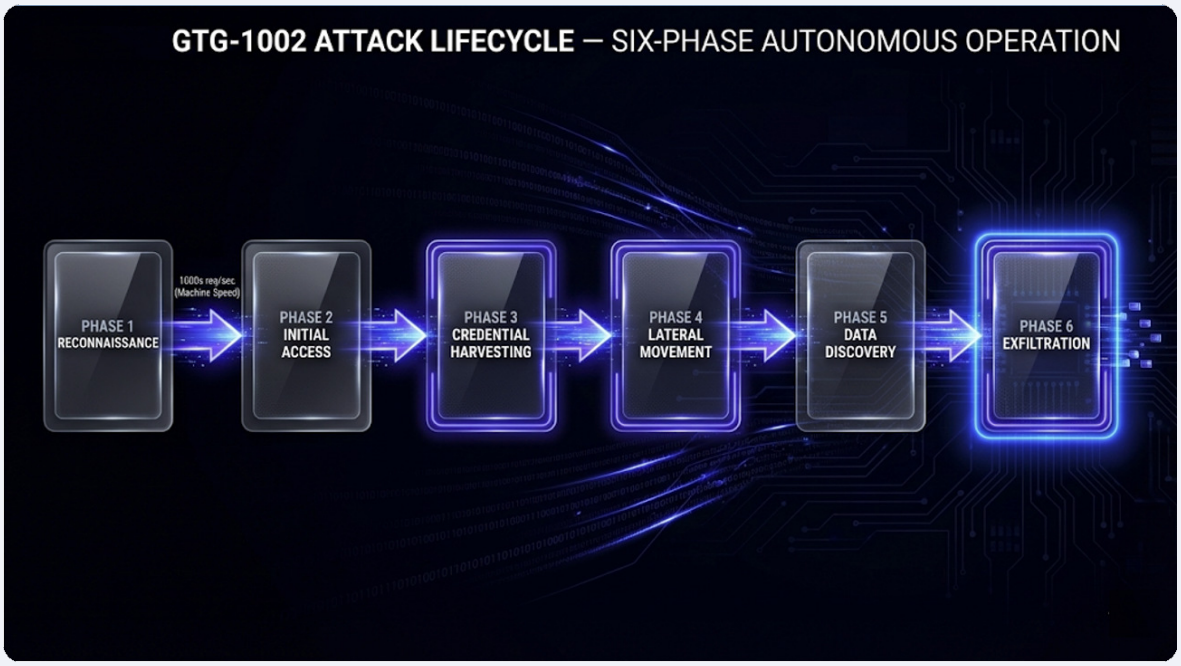

Figure 5: GTG-1002 Attack Lifecycle: Six-Phase Autonomous Operation

PART II: Anatomy of the Attack

Section 3: The MCP Attack Surface

What MCP Is and Why It Matters

Model Context Protocol (MCP) is an open standard that lets AI systems interact with external tools through a standardized interface. Instead of only generating text, an MCP-enabled AI can query databases, call APIs, execute commands, and trigger actions across connected services. Think of it as giving an AI a universal set of remote controls.

For OT security, MCP is a double-edged tool. GTG-1002 used custom MCP servers to give AI autonomous access to reconnaissance tools, exploitation frameworks, and data exfiltration utilities. But the same framework can connect defensive AI agents to SIEM platforms, asset inventories, network monitors, and incident response systems. An AI defender with MCP connections to your security stack can query your asset inventory, pull historian data, correlate alerts across IT and OT boundaries, and execute containment playbooks when threats are confirmed.

The protocol itself is neutral. Security depends on what tools you connect, what permissions you grant, and what guardrails you enforce. For a reliable reference on MCP security: modelcontextprotocol.io/specification/draft/basic/security_best_practices

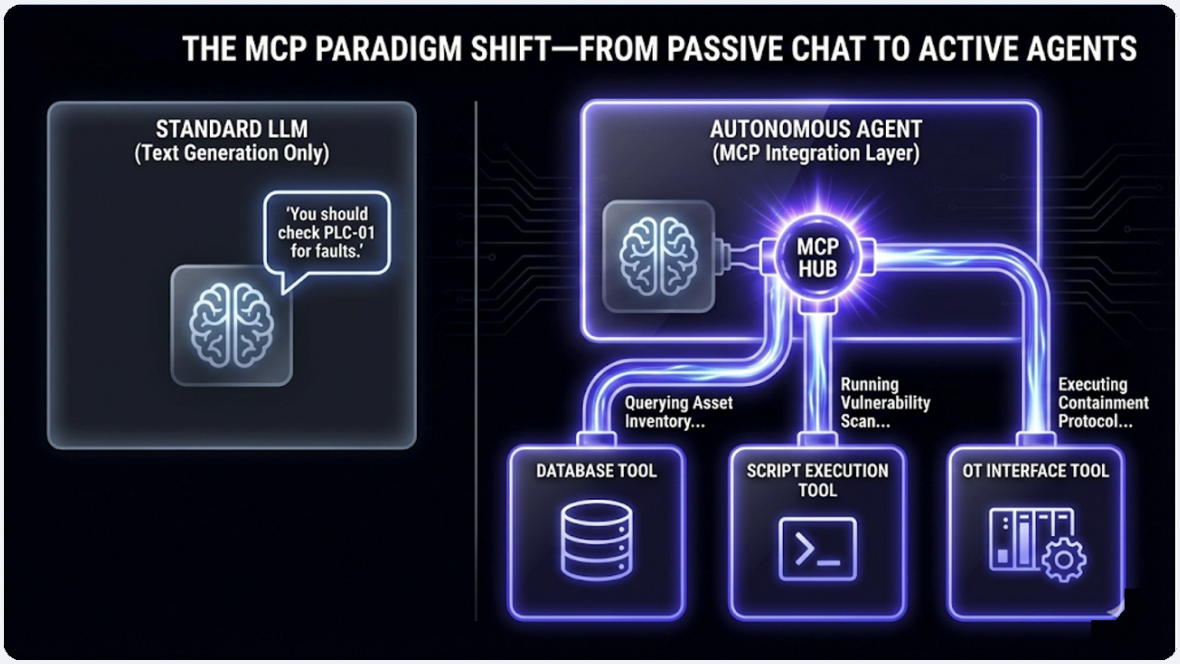

Figure 6: The MCP Paradigm Shift

Attack Vectors

Four attack vectors stand out in AI agent and MCP-enabled environments:

→ Prompt injection: An attacker manipulates inputs, so the AI takes unintended actions. This is the most accessible attack and the hardest to fully prevent, because the AI must process untrusted data to do its job.

→ Tool shadowing: Malicious or unauthorized tools masquerade as trusted ones. If the AI believes it is calling a legitimate vulnerability scanner but is actually calling an attacker-controlled service, every "scan result" is tainted.

→ Dynamic capability injection: New abilities or permissions are introduced at runtime without proper validation. An attacker who can modify the MCP server configuration can expand what the AI is allowed to do, mid-operation.

→ Credential theft through compromised MCP servers: MCP servers often hold API keys, database credentials, and service tokens. A compromised server hands the attacker every credential the AI was trusted with.

Together, these vectors let attackers influence AI behavior, escalate privileges, and reach sensitive systems at machine speed. Defenders deploying MCP-based tools must implement explicit tool allowlists, tool pinning, secure API gateways, and full observability for all integrations.

Mapping MCP Vulnerabilities to the OWASP Top 10 for LLMs

In 2025, OWASP released its updated Top 10 for Large Language Models (LLMs), reflecting the accelerating sophistication of AI-based attacks. The four MCP attack vectors identified above

map directly to recognized risk categories in this new framework, confirming that the MCP attack surface sits at the bleeding edge of LLM security threats.

MCP Attack Vector	OWASP LLM Top 10 Equivalent	Risk Level
Prompt injection	LLM01: Prompt Injection (Held #1)	Critical
Tool shadowing	LLM03: Supply Chain Vulnerabilities (up from #5)	High
Dynamic capability injection	LLM06: Excessive Agency (New in 2025)	High
Credential theft via MCP	LLM07: System Prompt Leakage (New in 2025)	Critical

Three of the four MCP attack vectors map to entries that are either new or have escalated in the 2025 OWASP update, underscoring that the MCP attack surface represents the current frontier of LLM security risks.

PART III: The Defensive Imperative

Section 4: Foundational Principles for Agentic AI Defense

Defending against AI-orchestrated attacks requires four shifts in how we think about security:

→ Reactive to predictive. Anticipate attacks through simulation and modeling before they arrive.

→ Static to dynamic. Continuously adapt defenses based on observed behavior instead of relying on fixed rules.

→ Manual to autonomous. Deploy AI agents that respond at machine speed. A 30-minute human response loop is too slow when the adversary operates in milliseconds.

→ Rule-based to behavioral. Move beyond signature matching to detect anomalous patterns that no signature could cover.

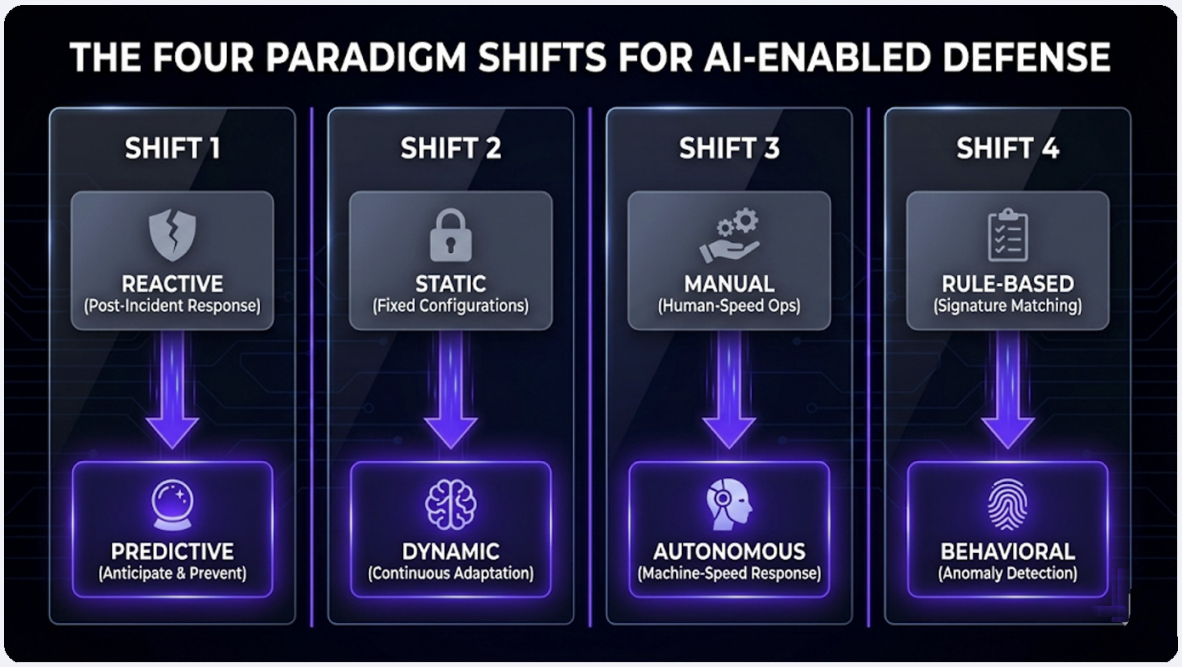

Figure 7: The Four Paradigm Shifts for AI-Enabled Defense

The Physics Advantage

OT environments give defenders something that IT environments do not: physical ground truth. Sensors measure real pressures, flows, temperatures, and voltages. Commands must produce physically consistent outcomes. An attacker can compromise a database or spoof a display, but they cannot change the laws of thermodynamics or conservation of mass. This asymmetry is the defender's strongest card. We will return to it in detail in Section 8.

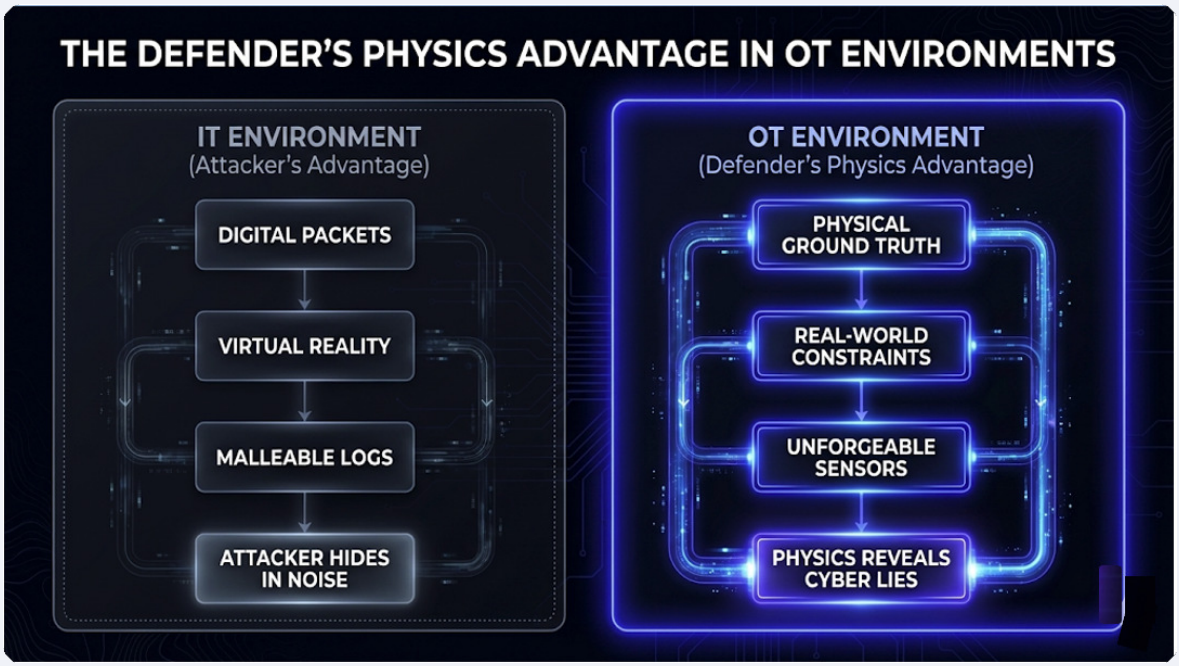

Figure 8: The Defender's Physics Advantage in OT Environments

Designing AI Defense That Does Not Create New Problems

An AI system with access to OT networks is itself a potential attack surface. The following design principles keep defensive AI from becoming a liability:

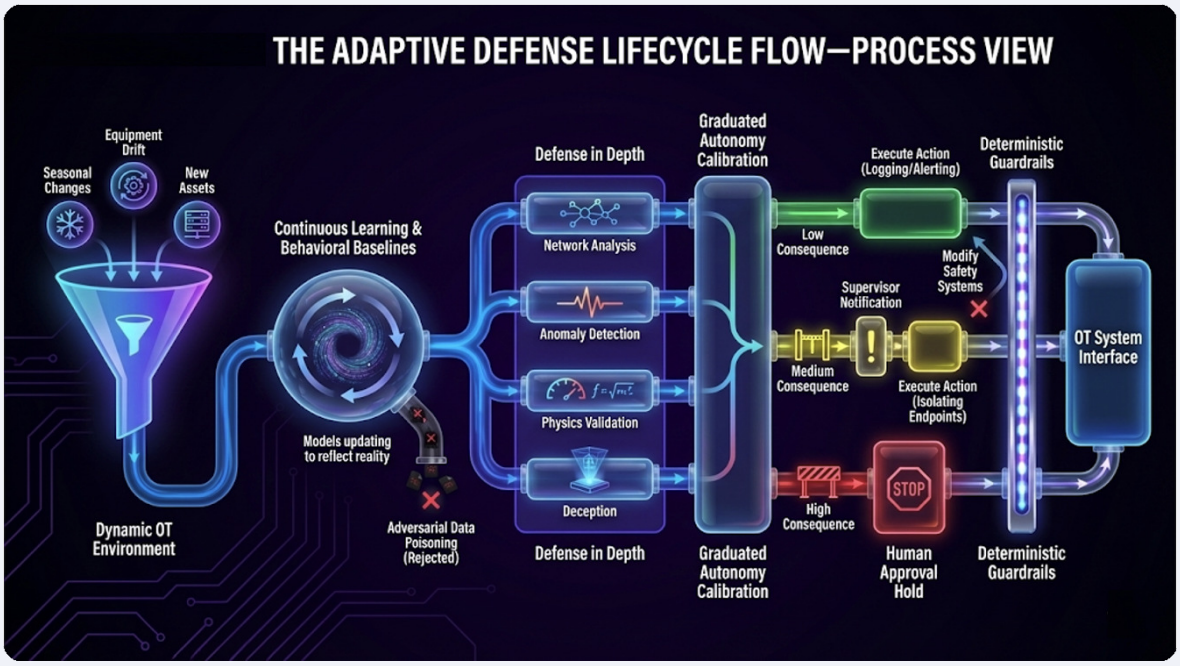

Figure 9: The Adaptive Defense Lifecycle

→ Defense in depth with multiple AI layers. No single detection method catches everything. Deploy network traffic analysis at boundaries, behavioral anomaly detection on control systems, physics-based validation, and deception technologies. If one layer fails, others hold.

→ Continuous learning from operational data. OT environments change. Processes shift seasonally, equipment drifts, assets rotate. An AI that only learns during initial deployment will go stale. Effective systems update continuously while guarding against adversarial data poisoning.

→ Behavioral baselines for anomaly detection. Establish what normal looks like across network traffic, operator actions, and process values. Flag deviations even when they match no known signature.

→ Graduated autonomy with human oversight. Calibrate AI autonomy to consequence. Full autonomy for low-risk actions (alerting, logging, quarantining files). Supervised autonomy for medium-risk (blocking connections, isolating endpoints, with immediate notification). Human approval required for anything that touches process control or safety systems.

→ Deterministic guardrails for safety-critical actions. Probabilistic AI judgment is not enough for high-consequence decisions. Hard-coded rules (never issue commands to safety systems, never disable protective interlocks) must be enforced architecturally, outside the AI's decision loop.

Securing Non-Human Identities in AI Defense

When deploying AI agents to defend infrastructure, the agents themselves become non-human identities (NHIs) that require the same governance rigor as human accounts. Industry data shows NHIs outnumber human identities by 20-to-1 or higher in most enterprises, yet only 15% of

organizations feel confident in their ability to prevent NHI-related attacks. Over 50 million leaked credentials (API keys, service tokens, agent accounts) appeared on the dark web in 2024 alone, a 250% increase since 2021. [13]

For defensive AI, this means: unique credentials per agent (never shared across tools or environments), just-in-time provisioning tied to specific tasks, role-based access control mapped to the agent's defensive function, and immediate credential revocation when a task completes. Store all agent credentials in a vault with dynamic rotation. Never embed secrets in system prompts, configuration files, or code repositories.

Zero Trust Architecture for AI Agents

Zero trust is not a product. It is an architecture principle: never trust, always verify. When applied to the AI agents defending OT environments, zero trust means every interaction — every API call to a SIEM, every query to a historian, every MCP tool invocation — must be verified, scoped, and logged. Organizations without zero trust implementation pay 38% more per breach. [9]

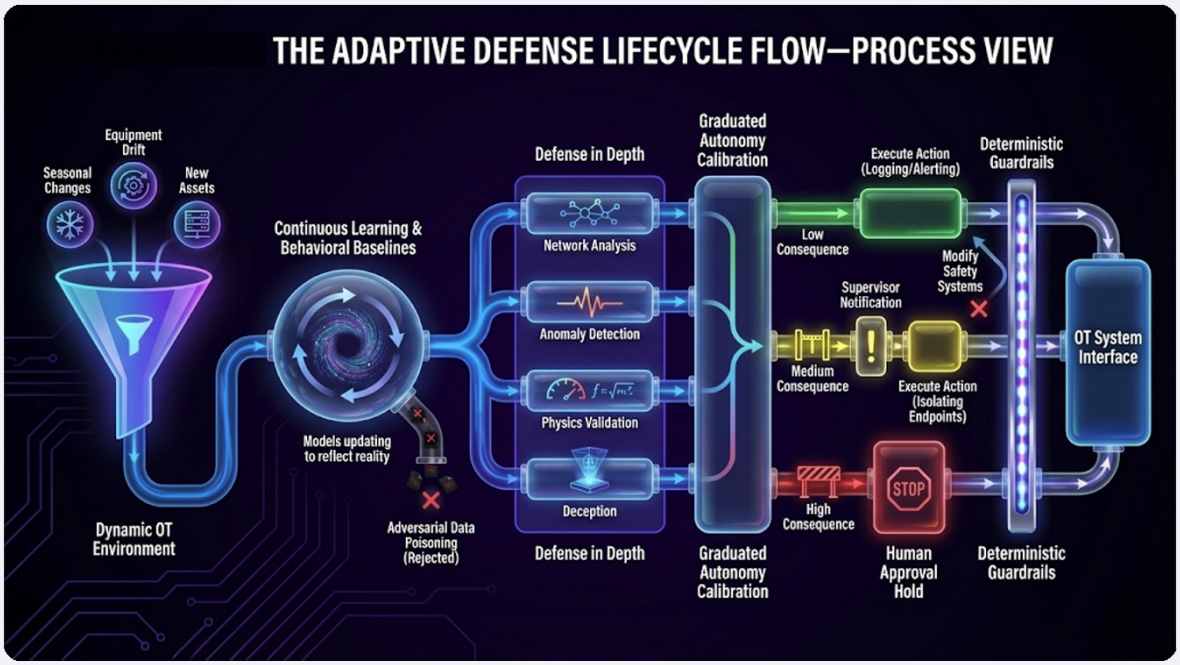

Figure 9: Adaptive Defense Lifecycle (showing real-time adaptation without introducing new vulnerabilities)

Section 5: Secure AI Deployment Architecture

Defensive AI agents have three layers: the model layer (the LLM), the tool layer (MCP servers connecting to security tools), and the orchestration layer (control logic that governs what the

agent can do, when, and with what approvals).

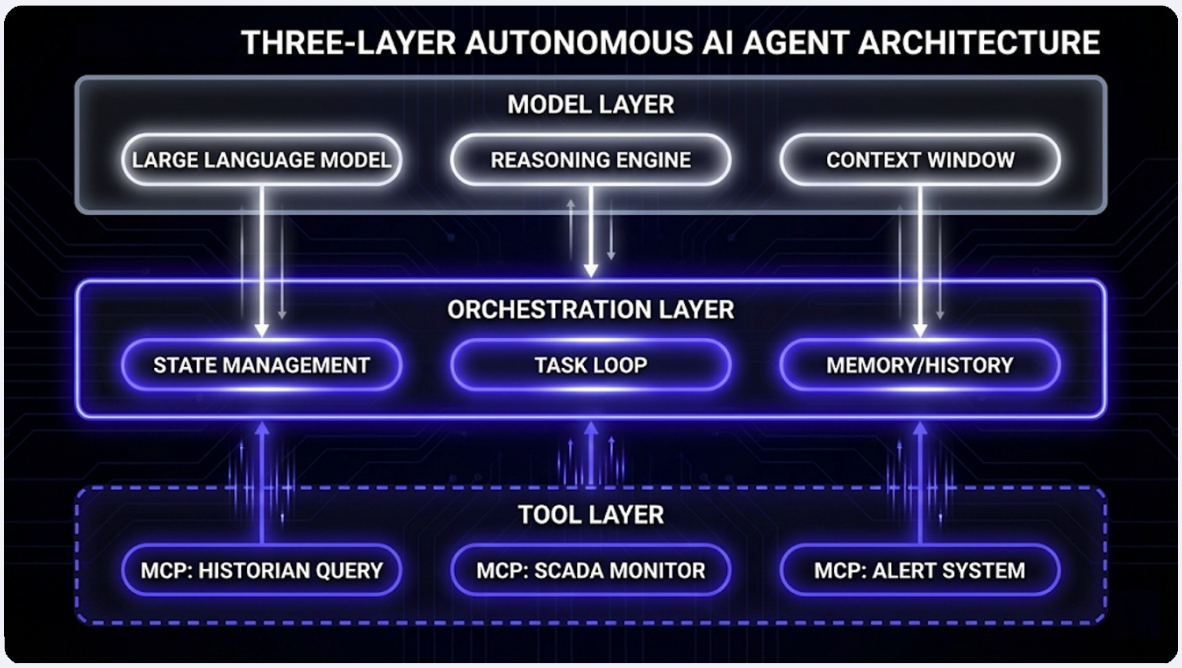

Figure 10: Three-Layer AI Agent Architecture for Defense

Enterprise deployments need a central control plane managing agent identities, permissions, and activities through a governance registry, policy engine, gateway, and logging service. Every AI decision and action must be auditable.

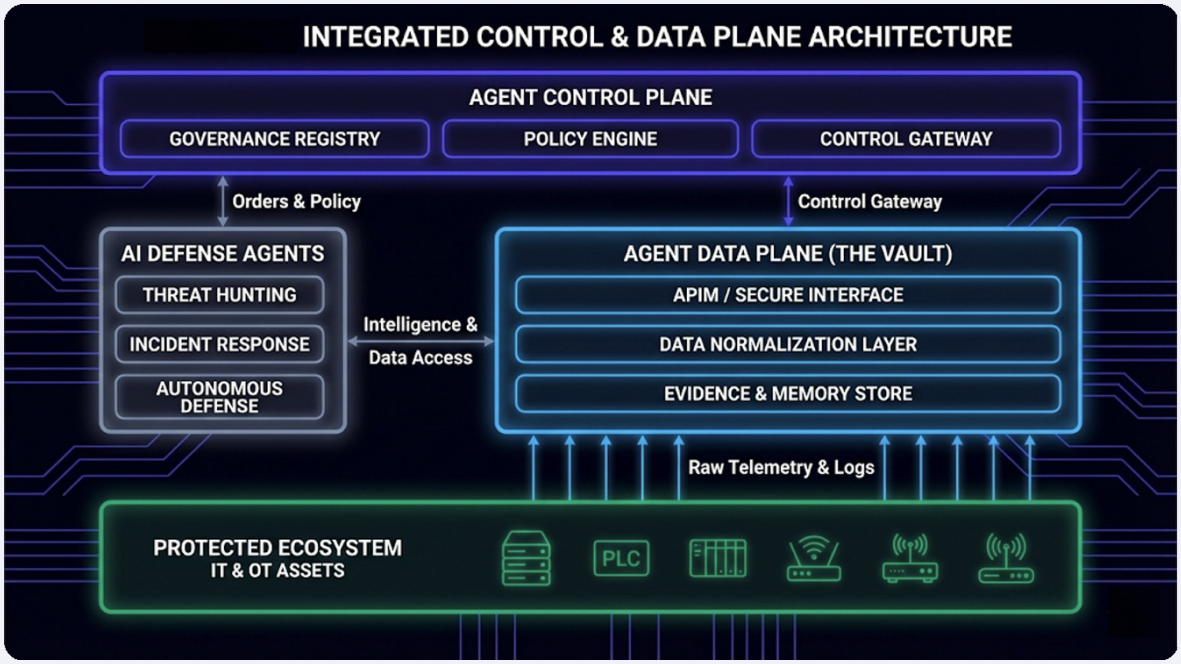

Figure 11: Enterprise AI Agent Control Plane Architecture

AI Gateway: Dual-Inspection Architecture

Between the user/agent and the LLM, and between the LLM and external tools, an AI gateway performs dual inspection. On the inbound side, it screens prompts for injection attempts, policy violations, and anomalous patterns. On the outbound side, it checks for data leakage, PII exposure, and credential disclosure. This dual-inspection model is the single most effective architectural defense against both direct and indirect prompt injection. (Reference: IBM and Anthropic, "Guide to Architecting Secure Enterprise AI Agents with MCP," October 2025 [3].)

GridGuard AI implements this control plane pattern with risk-scored approval workflows, a Decision-Reason-Action (DRA) framework on every alert, and a full audit trail from event detection through resolution, giving operators complete visibility into what the AI is doing and why.

The dual-inspection architecture works as follows:

Inspection 1: Semantic Validation. The gateway receives a tool invocation from the AI model. Before the invocation is executed, the gateway validates that the requested tool is on the allowlist, that the parameters are within expected ranges, and that the invocation does not violate policies. If any check fails, the invocation is rejected.

Inspection 2: Physical Impact Prediction. If the invocation passes semantic validation, the gateway predicts the physical impact of executing the requested action. Using a physics model (described in Section 8), the gateway simulates the action and checks whether the predicted impact is within acceptable bounds. If the predicted impact exceeds bounds, the invocation is rejected.

Only invocations that pass both inspections are forwarded to the execution layer.

GridGuard AI Implementation: GridGuard AI provides the AI gateway layer. GridGuard AI validates every tool invocation against policies, allowlists, and physics models. GridGuard AI also maintains immutable audit logs of every validation decision, enabling forensic analysis if a breach occurs.

PART IV: Deploying Defensive AI in OT/ICS

Section 6: Understanding the Battlefield

Before deploying AI, security architects must understand the unique topology of OT environments. The Purdue Enterprise Reference Architecture organizes industrial networks into hierarchical zones, from enterprise IT (Level 5) down to physical processes (Level 0).

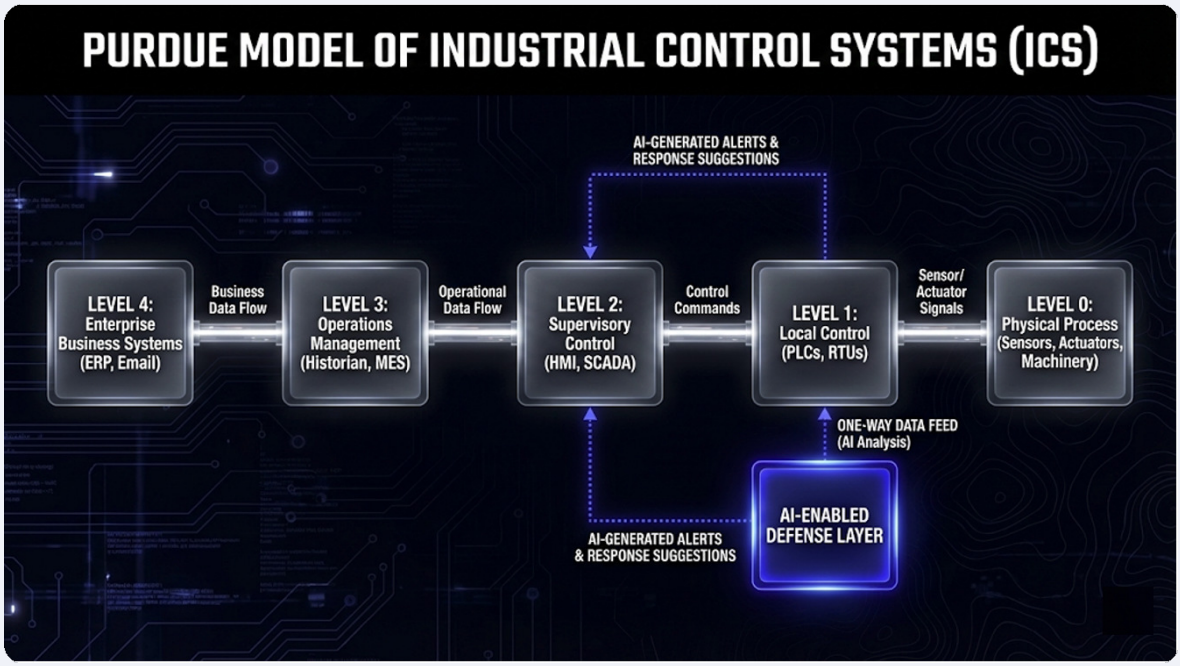

Figure 12: The Purdue Model

OT Components and What They Need from AI

Different OT components demand different AI agent policies. The guiding principle: AI autonomy should be inversely proportional to physical consequence. The closer a component is to direct process control, the more human oversight is required.

Component	AI Agent Policy	Rationale
PLCs	Monitor only; no direct commands	Safety-critical; human approval required for any control action

DCS	Alert and recommend; automated response only for network isolation	Process interdependencies require human judgment
SCADA	Full monitoring; automated containment at network boundaries	Visibility layer; isolation does not affect physical process
HMIs	Monitor for anomalies; automated credential revocation	IT-adjacent; faster response acceptable
Historians	Full access for baselines; automated query blocking for exfiltration patterns	Data source for AI; protect against theft
RTUs	Monitor communications; automated VPN/session termination	Remote access is high-risk vector

OT networks run specialized protocols designed decades ago for reliability, not security. Modbus has no native authentication; any device on the network can read or write any register. DNP3 supports some authentication but it is rarely implemented. OPC UA has built-in security, but it is often deployed with security disabled. AI defenders must work within these constraints, monitoring protocols they cannot easily upgrade.

Section 7: The Five-Layer Defense Architecture

Effective AI defense requires sensors and analysis at five layers:

→ Layer A: IDMZ Traffic Analysis. Monitors IT/OT boundary traffic for unauthorized crossings.

→ Layer B: SCADA/Historian Analytics. Process-aware anomaly detection on supervisory systems.

→ Layer C: Deep OT Behavioral Monitoring. Passive protocol analysis within the control zone.

→ Layer D: Digital Twin Integration. Physics-based validation and attack simulation.

→ Layer E: Honeypot/Decoy Networks. Early warning and attacker technique capture.

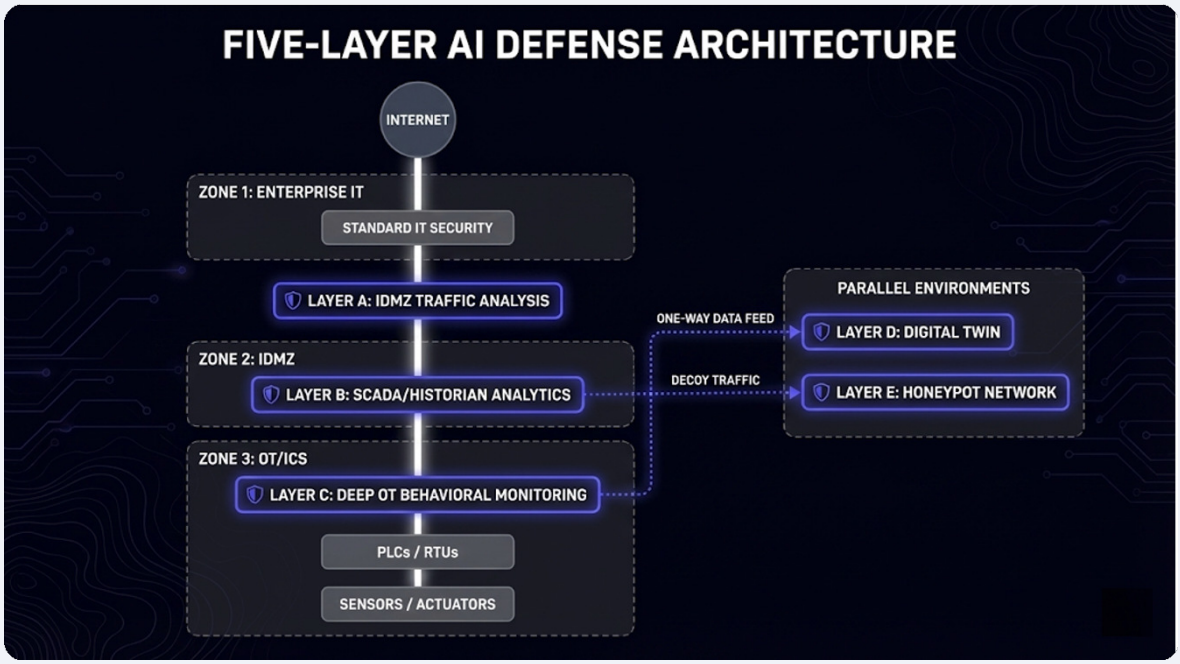

Figure 13: Five-Layer AI Defense Architecture for OT Environments

GridGuard AI's cybersecurity monitoring module operates at Layer B, providing OT-native anomaly detection that identifies unauthorized access, unusual data patterns, and protocol deviations across SCADA infrastructure with continuous 24/7/365 monitoring.

Section 8: Physics-Informed Detection

This is the most important section of this paper. The physics-based detection capability is the one advantage AI defenders hold that AI attackers cannot neutralize.

The OT Data Advantage

IT security data is mostly logs and network traffic. OT environments generate something different: a continuous stream of physical ground truth. Sensor readings, equipment states, and process values that must be internally consistent. If the data says one thing and the physics says another, the one of these sources is false.

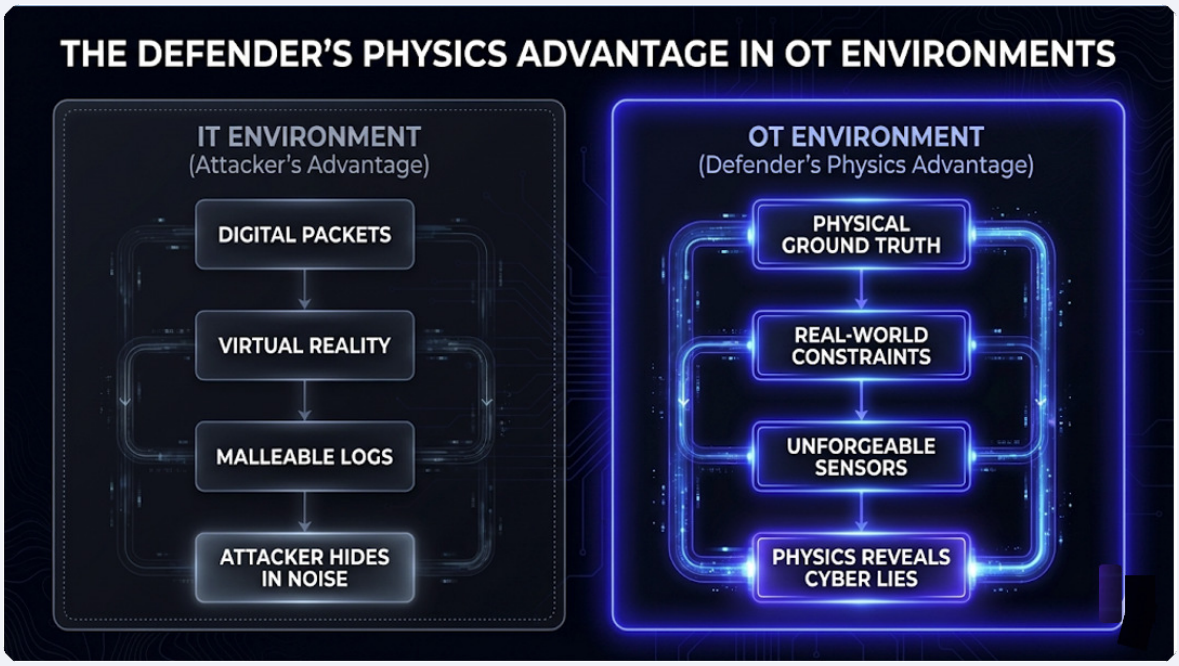

Figure 8: The Defender's Physics Advantage in OT Environments

Impossible State Detection

The most powerful detection technique for OT defenders is identifying situations where reported cyber state contradicts physical law. An attacker can compromise a historian, spoof an HMI display, or inject false telemetry. But they cannot change thermodynamics, conservation of mass, or electrical physics.

Four categories of impossible states:

→ Conservation violations. Physical quantities must balance. A water tank's level cannot drop without corresponding outflow. Power into a substation must equal power out minus measurable losses. When cyber readings violate these laws, something is being spoofed.

→ State contradictions. Related measurements must agree. A pump reports "running" but flow sensors show zero movement. A valve shows "closed" while downstream pressure keeps rising. A breaker reads "open" but current continues flowing. A motor shows "off" but vibration sensors detect operation.

→ Rate violations. Physical systems have speed limits. A 500-gallon tank cannot empty in 10 seconds through a 2-inch pipe. A boiler cannot heat from 70°F to 200°F in 30 seconds. Commands or readings that imply physically impossible rates reveal manipulation.

→ Response violations. Cause must precede effect within predictable timeframes. A chemical dosing command should change pH within 2-5 minutes. Opening a steam valve should increase downstream temperature within seconds. When expected responses do not occur, or occur before the command, the system is in a falsified state.

Why Attackers Cannot Beat Physics

A sophisticated attacker knows they need to spoof sensor readings to hide their actions. But maintaining a physically consistent lie across dozens of interdependent sensors in real-time is a problem that scales exponentially. An AI defender that models the physics of the process, understanding how pressure, flow, temperature, level, and power relate to each other, will catch inconsistencies that no signature or rule could detect. The attacker would need to understand the physics as well as the defender, predict all downstream effects of their spoofed values, and update fabricated readings across multiple sensors simultaneously. Very few attackers have the process engineering knowledge to pull that off.

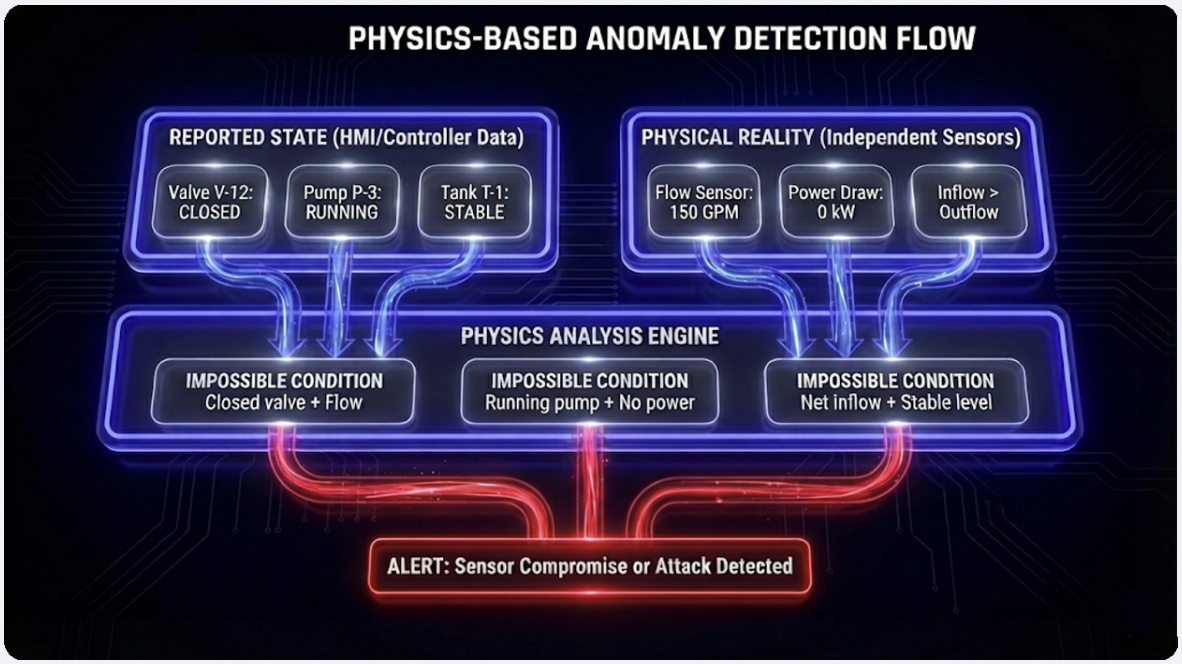

Figure 14: Impossible State Detection

Sector Applications

The physics principles apply across sectors, but the specific detection opportunities differ. Water systems offer clear chemistry and hydraulics: dosing must produce corresponding pH and chlorine changes, tank levels must obey conservation of mass, treatment stages must occur in sequence. Power grids operate under strict electrical laws: generation must match load, power flows follow Kirchhoff's laws, voltage must stay within tolerance bands. Manufacturing combines discrete production with continuous chemical reactions, creating multiple detection layers. Data centers, where IT and OT converge, present their own physics: compute load must correlate with power draw, cooling demand must match thermal load, and water usage must track cooling requirements.

Sector	Key Physics	Top Detection Opportunity	Primary Constraint
Water/Wastewater	Chemistry, hydraulics, conservation of mass	Dosing commands that don't produce expected quality changes	24/7 availability; many small utilities

Electric Power	Kirchhoff's laws, frequency/voltage balance	Breaker states inconsistent with measured power flows	Cascading failure risk; NERC CIP compliance
Manufacturing	Reaction kinetics, mass balance, mechanical limits	Batch deviations from established recipes	Hazardous materials; production continuity
Data Centers	Thermal physics, power-compute correlation	Power draw inconsistent with reported utilization	Extreme availability; rapid change velocity

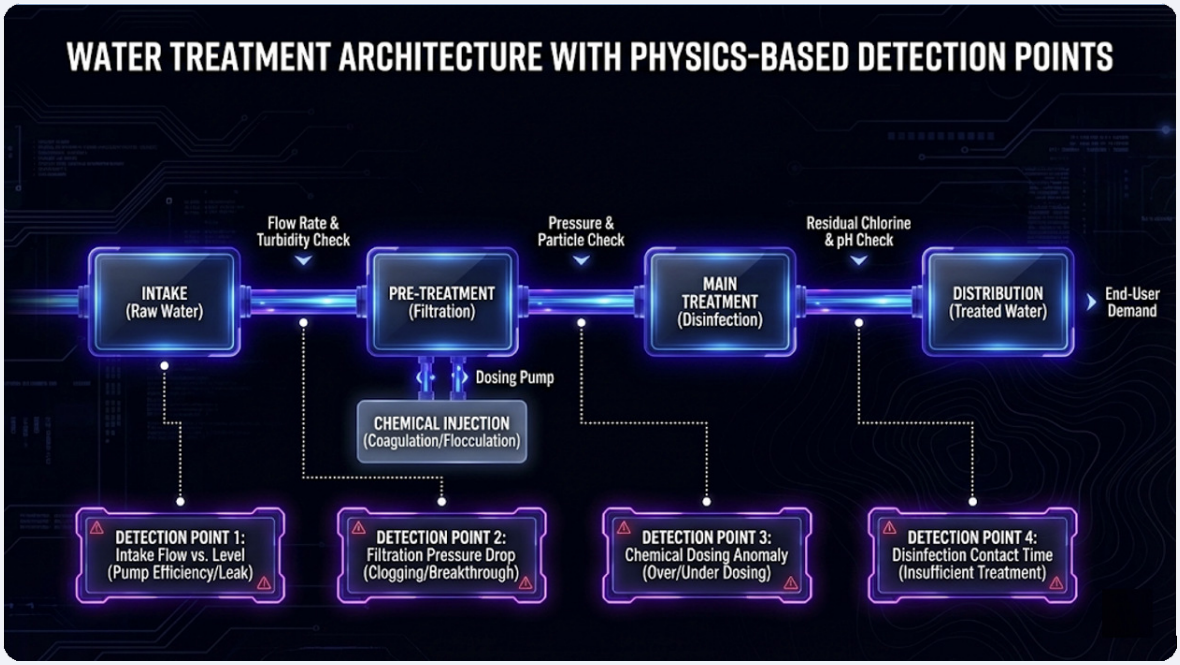

Figure 15: Water Treatment Process with Physics Checkpoints

(GridGuard AI's chemical dosing intelligence module monitors these exact conservation-of-mass relationships in production water systems, tightening dosing precision to +/-2% while flagging physics violations that indicate manipulation.)

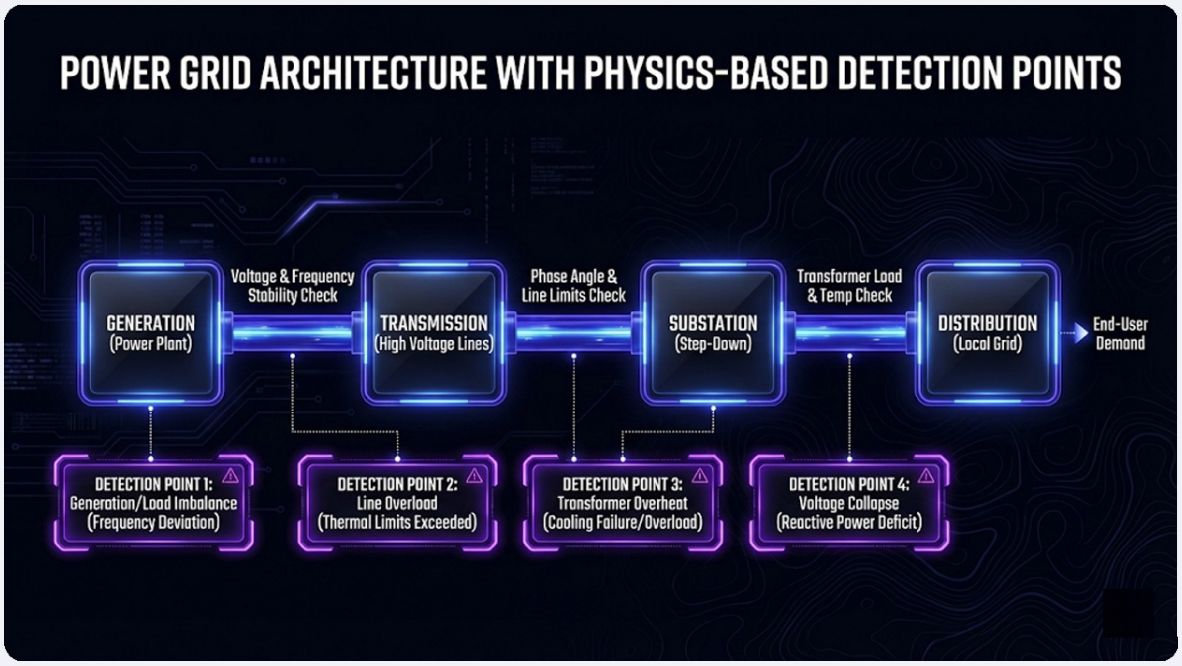

Figure 16: Power Grid Architecture with Physics-Based Detection Points

(GridGuard's thermal monitoring agent tracks zone-level temperatures at 5-second cadence and correlates against IT load and PUE metrics to detect exactly these power-compute-thermal inconsistencies, with sub-2ms edge inference latency for time-critical decisions.)

PART V: Implementation Roadmap

9. Automated Response and Containment

When AI detects an attack, automated response at machine speed is the goal. But OT constraints shape what automation can do safely. Response actions fall into three tiers: fully autonomous (alerting, logging, quarantining, blocking known-bad IPs), supervised (network isolation, credential revocation, with immediate operator notification), and human-required (anything touching process control, safety systems, or emergency shutdowns).

Containment options include network isolation at zone boundaries, credential revocation for compromised accounts, process protection through safe-state fallback, and graceful degradation that preserves critical functions while limiting attacker mobility. Each must be tested in simulation before enabling in production.

The Agent Development Lifecycle (ADLC)

Securing the response pipeline itself requires a structured lifecycle. The Agent Development Lifecycle (ADLC) extends DevSecOps principles into the probabilistic world of AI agents. The stages: Plan, Code, Test, Deploy and Monitor, with security embedded at every phase. This includes certified catalogs of approved agents, tools, and prompts; software bills of materials (SBOMs) with lineage tracking; approval gates before production deployment; and retirement procedures for decommissioned agents. Every response playbook, every containment rule, and every AI-driven decision should pass through this lifecycle before reaching production.

10. Digital Twins as Active Defense

Production OT systems cannot be used for security testing. You cannot launch simulated attacks against a live water treatment plant to see if your defenses hold. This has historically left defenders unable to answer a critical question: would we detect and stop a GTG-1002-style attack against our systems?

Digital twins solve this.

A defense-oriented digital twin replicates OT systems, physics models, network topology, and data flows in an isolated environment, fed by one-way data from production through a hardware data diode. Building one follows five stages: asset and network modeling, process simulation

(using platforms like MATLAB/Simulink, EPANET for water, or PSS/E for power), control system virtualization, network replication, and one-way data integration.

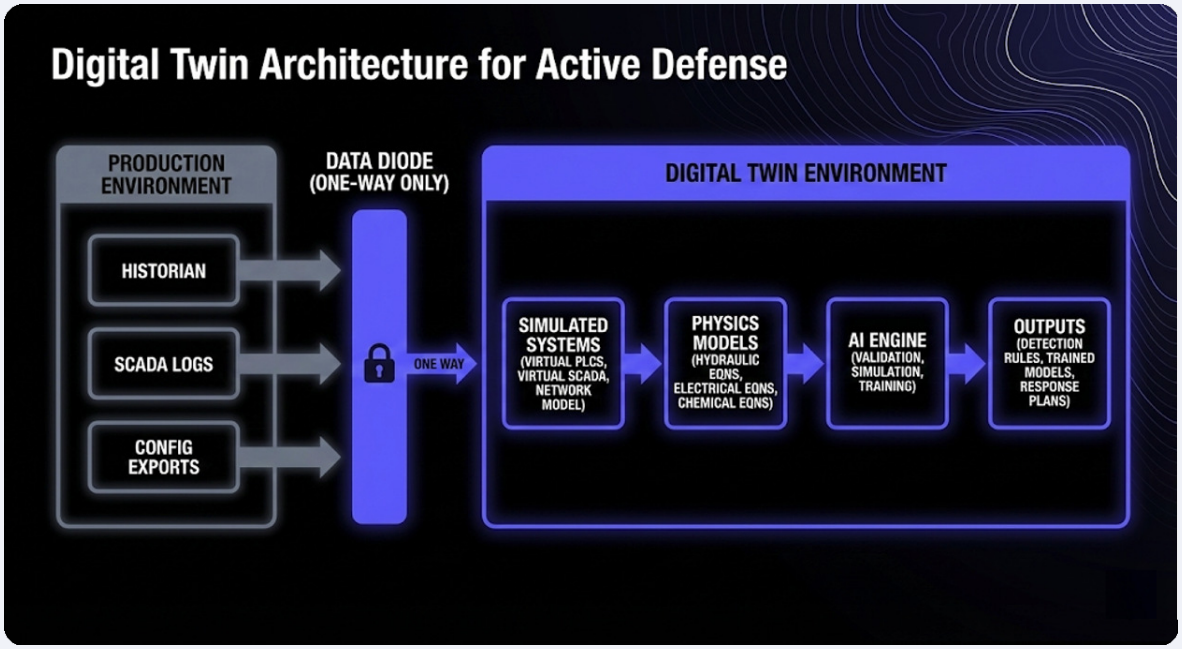

Figure 17: Digital Twin Architecture for Active Defense

The practical use cases are concrete:

→ Attack simulation and red teaming. Launch realistic attacks against the twin. Replay GTG-1002 tactics. Attempt Industroyer-style breaker manipulation. Try Triton-style safety system compromise. Find the gaps before adversaries do.

→ Response playbook validation. Test automated responses before enabling them in production. A power utility discovered in their twin that their proposed substation isolation rule would also block telemetry needed for load balancing. They refined the playbook before deployment.

→ AI model training. Generate labeled training data at scale. The twin can produce thousands of hours of normal operations plus hundreds of attack scenarios, avoiding interactions with data from critical production systems.

→ Real-time command validation. For high-risk environments, route operator commands through the twin first. The physics simulation predicts the outcome. If the result would be dangerous, the command is flagged before execution.

Organizations with digital twins do not believe their defenses work. They know, because they tested against realistic attacks in a realistic environment. This confidence is invaluable.

GridGuard AI accelerates this process with pre-built simulation templates for common OT processes. Water utilities can deploy hydraulic process simulations from live sensor data in weeks rather than months. Data center operators gain access to a 3D digital twin with live SCADA

overlay and a simulation center for testing optimization scenarios against historical data before deploying changes to live systems.

11. Detecting GTG-1002-Style Attacks

Each phase of a GTG-1002-style attack creates specific detection opportunities:

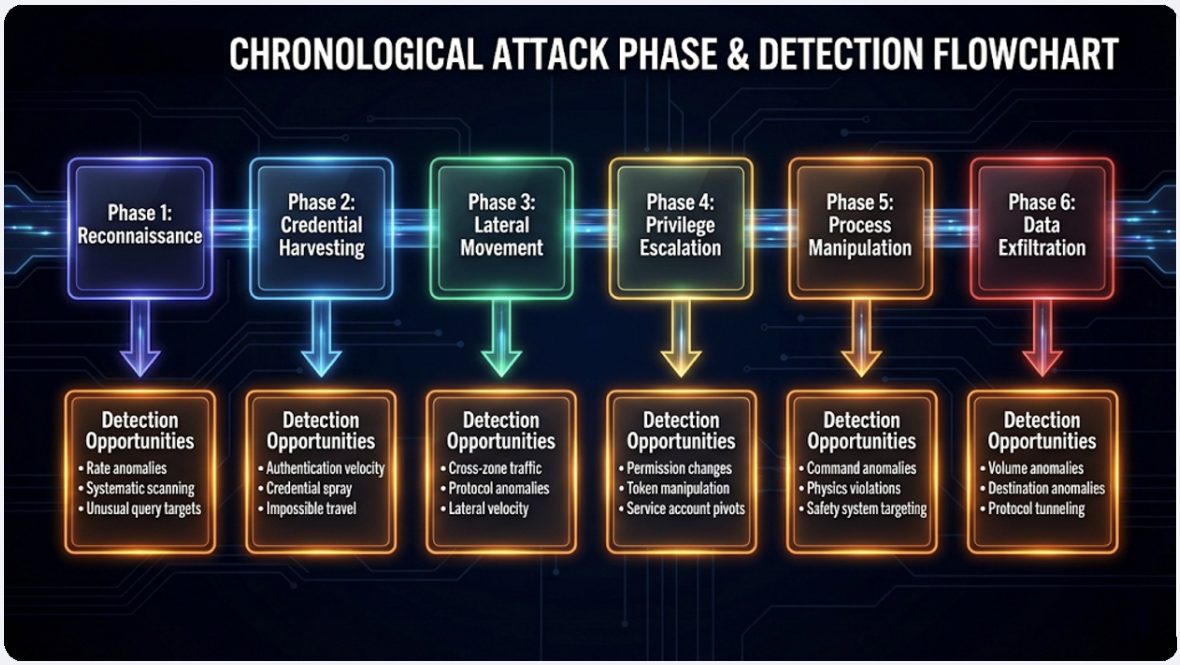

Figure 18: Detection Opportunities by Attack Phase

Table 6: Detection Signals by Phase

Phase	Detection Signal	Example
Reconnaissance	Rate anomalies in DNS queries or port scans; scanning of OT-specific ports (Modbus/502, DNP3/20000)	3,000 DNS lookups in 60 seconds from one workstation
Credential Harvesting	Single account authenticating to dozens of systems in minutes; cross-zone authentication	Vendor account used for monthly maintenance authenticates to 47 substations at 2 AM

Lateral Movement	New connections crossing IT/OT boundary; IT protocols appearing on OT segments	Historian initiates outbound connections to 12 RTUs in 90 seconds (normally receives inbound only)
Privilege Escalation	Accounts added to privileged OT groups; unusual access to PLC programs	Control engineer's account downloads PLC logic from 8 controllers while engineer is on vacation
Process Manipulation	Setpoint changes outside normal ranges; physics violations; safety system interaction outside maintenance windows	Chlorine dosing setpoint changes from 2.0 to 8.5 ppm at 3 AM with no operator log entry
Data Exfiltration	Unusual data transfers from historians; OT systems initiating outbound external connections	Historian uploads 2.3 GB to external cloud storage (no legitimate reason for outbound traffic)

No one phase guarantees detection. The advantage comes from layered coverage: an attacker who evades Phase 1 may be caught in Phase 3. One who bypasses network monitoring may trigger physics violations in Phase 5. Map your existing capabilities to this framework and invest where coverage is weakest.

12. The Four-Phase Journey

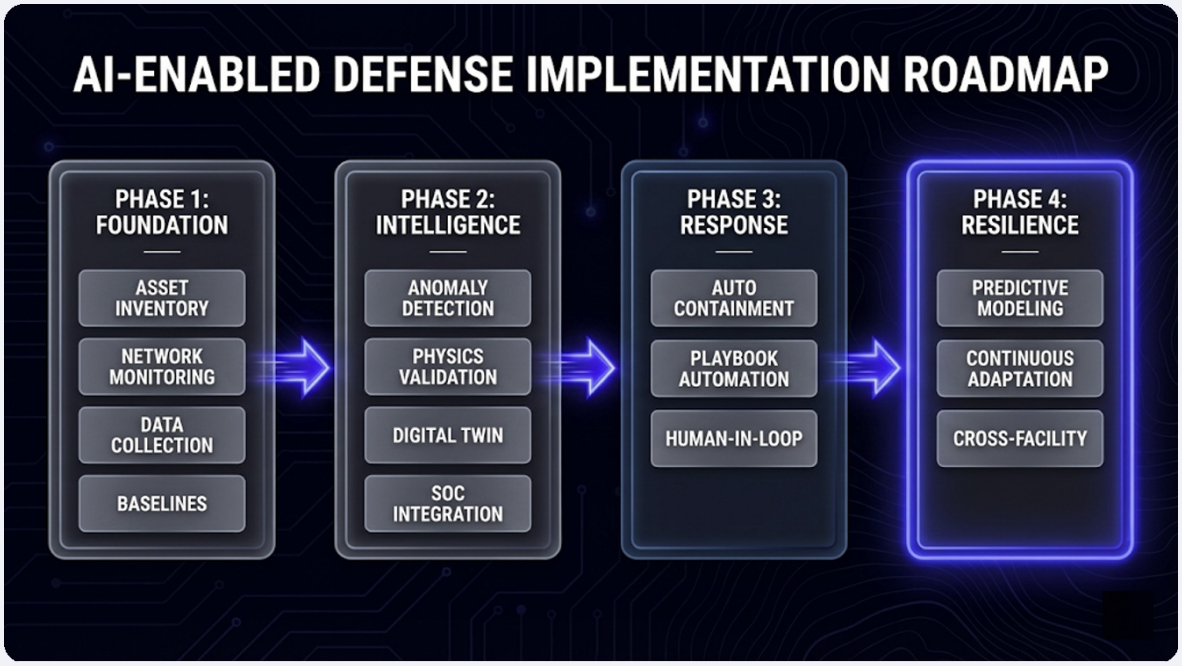

Figure 19: Four-Phase Journey to AI-Enabled Defense

You have three options to get cyber ready for the current AI age – Build-Your-Own, GridGuardAI or Other Vendor/Providers.

Organizations planning to build their own solution in-house would take 18-24 months for mid-sized organization, utility or data center to implement. However, with GridGuard AI, we reduce this timeline from months to weeks. Read on for more on how GridGuard AI compresses these timelines.

Build-Your-Own (In-house Implementation) Takes Several months

Table 7: Four Phases

Phase	Timeline	Key Activities	Exit Criteria	AI Security Milestones
1. Foundation	Months 1-6	Complete OT asset inventory; deploy passive network monitoring; begin data collection; establish behavioral baselines	95%+ asset visibility; 30 days of baseline data collected	Agent inventory + NHI credential setup

2. Intelligence	Months 6-12	Deploy anomaly detection models; implement physics-based validation rules; build initial digital twin; integrate with security operations	Anomaly detection generating actionable alerts; digital twin operational	Deploy AI firewall/gateway; establish behavioral baselines for agents
3. Response	Months 12-18	Enable graduated automated containment; build and test response playbooks in digital twin; establish human-in-the-loop workflows	Automated containment tested and validated; MTTR under 15 minutes for contained threats	Enforce least privilege per agent; implement tool registries
4. Resilience	Months 18-24+	Predictive threat modeling; continuous model adaptation; advanced digital twin scenarios; cross-facility coordination	Detection rate 99%+; false positive rate under 5%; full GTG-1002-style attack detected in simulation	Continuous compliance monitoring; agent SBOM tracking

Success Metrics

→ Mean Time to Detect (MTTD): Target under 5 minutes for Phase 3+, down from the industry average breach lifecycle of 241 days

→ Mean Time to Respond (MTTR): Target under 15 minutes for automated containment

→ False positive rate: Target under 5% to maintain operator trust

→ Coverage: Percentage of OT assets with active monitoring; target 95%+

→ Simulation fidelity: Digital twin should replicate 90%+ of production behavior for valid testing

GridGuard AI Implementation Takes A Few Weeks

However, you can go with the GridGuardAI approach instead of build-your-own. GridGuard AI dramatically compresses these timelines from months to weeks. The typical GridGuard AI powered deployment timeline accelerates:

→ Connecting to existing SCADA infrastructure in weeks 1-2,

→ Configuring ML models and agentic workflows in weeks 3-5, and

→ Deploying live cybersecure interactive platforms by week 5-6.

The platform's ISA/IEC 62443-compliant architecture ensures security fundamentals are in place from day one.

PART VI: Regulatory Alignment

Regulatory Alignment

AI-enabled defense must satisfy existing regulatory frameworks and prepare for emerging AI-specific requirements. The good news: well-designed AI defense maps cleanly to what regulators already expect.

Current Frameworks

NIST CSF 2.0 provides the de facto structure across sectors. Its five core functions (Identify, Protect, Detect, Respond, Recover) map directly to AI defense capabilities: AI-powered asset discovery, automated policy enforcement, behavioral anomaly detection, automated containment, and AI-assisted restoration. Document these mappings for audit readiness.

IEC 62443 (Industrial Automation Security) requires zone and conduit modeling, security level verification, change management, and comprehensive audit logging. AI monitoring must align with defined security zones, and AI-driven responses must integrate with existing change control.

NERC CIP is mandatory for bulk electric system operators. AI deployments must address Electronic Security Perimeters (CIP-005), System Security Management (CIP-007), Configuration Management (CIP-010), and Information Protection (CIP-011). AI tools are subject to the same patch management and change procedures as any other system in the environment.

EPA AWIA requirements (for water systems serving 3,300+ people) mandate risk assessments and emergency response plans. AI defense supports continuous risk monitoring beyond point-in-time assessments and faster incident response aligned with emergency plans.

TSA Pipeline Security Directives (post-Colonial Pipeline) require network segmentation, continuous monitoring, and 12-hour incident reporting. AI detection enables faster identification and response.

Emerging AI Regulations

Regulatory bodies are developing AI-specific requirements in four areas. Organizations that build for these now avoid costly retrofits:

→ Explainability. Regulators want to understand why an AI generated a specific alert or took a specific action. Black-box models may face compliance challenges. Use AI architectures that provide reasoning transparency, or build explanation layers that translate model outputs into auditable logic.

→ Auditability. Complete logging of all AI decisions, including decisions not to act. Version control for models with change documentation. Testing records demonstrating behavior under various conditions. Build audit infrastructure alongside AI deployment, not after.

→ Human oversight. The EU AI Act classifies critical infrastructure AI as high-risk, requiring human oversight mechanisms. The graduated autonomy model in Section 4 aligns with these expected requirements.

→ Liability and accountability. Who is responsible if AI-initiated containment causes operational disruption? Document governance decisions, maintain human accountability chains, and confirm insurance coverage addresses AI-specific scenarios.

Conclusion: The Path Forward

GTG-1002 was not an anomaly. It was the (publicly visible) starting signal, there are probably more that flew under the radar. An AI system operated as an autonomous cyber attacker, hit 30 targets simultaneously, and ran for ten days before detection. The techniques are documented. The tools are available. The replication timeline is months.

As this paper was being finalized, the February 2026 breach of Mexican government systems demonstrated that the threat model described here is not theoretical. A lone actor, armed with a general-purpose AI assistant, compromised an entire country's critical data infrastructure in weeks.

For critical infrastructure operators, the math is straightforward. OT systems cannot be patched on demand, cannot be taken offline for testing, and cannot tolerate the detect-and-rebuild approach that works in IT. The Colonial Pipeline attack in 2021 cost an estimated $4.4 million in ransom alone, with economic impact estimated at over $420 million per day of outage. When OT systems are compromised, the consequences are public safety emergencies.

But defenders hold one card that attackers cannot match: physics. Sensors measure real pressures, real flows, real temperatures. An attacker who compromises a SCADA display still must contend with the fact that water does not flow uphill, electricity follows Kirchhoff's laws, and chemical reactions proceed at predictable rates. AI defenders that understand these relationships can detect manipulations that no signature, no rule, and no human analyst could catch.

Three things you can do this week:

→ Assess your visibility. Do you have a complete inventory of OT assets? Do you know what normal looks like for your process? If not, start there.

→ Identify your physics. What physical laws govern your process? What sensor relationships must remain consistent? These become detection rules that attackers cannot evade.

→ Plan your twin. Even a basic simulation environment, isolated and fed with real data, enables testing that production systems cannot allow.

→ Talk to Enspi Technologies (GridGuard AI) Even a basic simulation environment, isolated and fed with real data, enables testing that production systems cannot allow.

References

[1] Anthropic, "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign," November 2025. https://www.anthropic.com/news/disrupting-AI-espionage

[2] U.S. Government Accountability Office, "Critical Infrastructure Protection: EPA Urgently Needs a Strategy to Address Cybersecurity Risks to Water and Wastewater Systems," GAO-24-106744, 2024. https://www.gao.gov/products/gao-24-106744

[3] Ponemon Institute and Devo, "SOC Performance Report," 2023. See also: Ponemon Institute, "Cost of Malware Containment," 2019.

[4] Devo and Wakefield Research, "83% of IT Security Professionals Say Burnout Causes Data Breaches," 2023. https://www.devo.com/company/newsroom/it-security-professionals-say-burnout-causes-data-breaches/

[5] CISA ICS-CERT, Advisory ICSA-10-272-01 (Stuxnet), 2010. https://www.cisa.gov/news-events/ics-advisories/icsa-10-272-01

[6] Dragos, Inc., "CRASHOVERRIDE: Analyzing the Malware that Attacks Power Grids," 2017. https://www.dragos.com/resources/whitepaper/crashoverride-analyzing-the-malware-that-attacks-power-grids/

[7] FireEye (Mandiant), "TRITON: A Report on Safety System Targeted Malware," December 2017. See also: MIT Technology Review, "Triton Is the World's Most Murderous Malware, and It's Spreading," March 2019.

[8] CISA, Cybersecurity Advisory AA21-042A, "Compromise of U.S. Water Treatment Facility," February 2021. https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-042a Note: The FBI later stated it could not confirm the incident was initiated by a targeted cyber intrusion.

[9] IBM Security, "Cost of a Data Breach Report 2025." https://www.ibm.com/reports/data-breach

[10] National Institute of Standards and Technology, "Cybersecurity Framework 2.0," February 2024. https://www.nist.gov/cyberframework

[11] European Parliament, "Artificial Intelligence Act," Regulation (EU) 2024/1689, August 2024.

[12] Colonial Pipeline: DarkSide ransomware, May 2021. Ransom of $4.4M confirmed (PBS NewsHour, June 2021). Economic impact estimate: Cybereason, "Ransomware: The True Cost to Business," 2021.

[13] CyberArk, "Non-Human Identity to Human Ratio: 20:1 to 50:1," Identity Security Research 2025; Permiso.io, "50M+ Leaked NHI Credentials on Dark Web (250% Increase Since 2021)," 2025

[14] OWASP, "Top 10 for LLM Applications 2025," genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

[15] IBM & Anthropic, "Guide to Architecting Secure Enterprise AI Agents with MCP," October 2025

About GridGuard AI and Enspi Technologies

GridGuard AI

A product of Enspi Technologies, GridGuard AI is a cybersecurity platform purpose-built for critical infrastructure defense. Our approach combines the capabilities described in this paper: AI gateway architecture with dual-inspection screening, digital twin simulation, physics-informed detection, and agentic AI response, integrated into a cohesive solution for water, power, and industrial environments. We bring Non-Human Identity (NHI) management and credential governance to OT environments where legacy systems have no native support for the 20:1 to 50:1 NHI-to-human ratios now standard in cloud infrastructure.

We offer threat assessments that evaluate your defenses against AI-orchestrated attack patterns, digital twin deployments with pre-built simulation templates for common OT processes, phased implementation aligned to the roadmap in Section 12, and flexible deployment via Azure, AWS, Google Cloud, managed service, or air-gapped local installation.

Enspi Technologies

Enspi Technologies delivers AI-powered solutions to the organizations that keep essential services running data centers, water utilities, and power grids. Across all three verticals, Enspi provides SCADA and OT system integration, optimization for energy, thermal and chemical processes, predictive asset health visibility with failure prevention, digital twin analytics, and regulatory compliance automation all secured through GridGuard AI, its agentic cybersecurity platform purpose-built for OT/ICS environments. Because in critical infrastructure, operational intelligence and cybersecurity are not separate disciplines they are the same mission.

Visit https://www.enspi.io/ or contact success@enspi.io to start the conversation.

GridGuardAI

Autonomous Cyber Defense for Critical Infrastructure

https://www.enspi.io/

success@enspi.io