NOZOMI NETWORKS

Valid Credentials, Native Tools, No Malware: Agents in Industrial Control Systems

One you invited. One you didn't. Why there is no kill switch, and what to build instead.

Chris Grove, Director of Cybersecurity Strategy Nozomi Networks

Taking the agenda literally

The agenda says a rogue AI agent may already be in your system. **This session takes that sentence literally.**

© 2026 Nozomi Networks Inc. | All rights reserved. | **nozominetworks.com**

Two kinds of agents, one definition

Both are already in your control network. Only one was on the purchase order.

The invited agent

Copilots and agentic assistants that automation vendors now embed in engineering suites, HMIs and controllers. They read project files, generate and modify logic, and increasingly act on your behalf, inside the trust your engineers already have.

The uninvited agent

Attackers chaining models and agents into intrusions that run at machine speed: reconnaissance, tooling, the IT-to-OT pivot and persistence, with a person approving only the few decisions that matter.

One working definition. An agent is anything that reads state, decides, and acts on your systems without a person typing each step. Whose payroll it is on does not change the definition.

PART 1

What AI changed, and what it didn't

PART 2

The agents you installed on purpose

PART 3

The kill-switch question, and a 90-day plan

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

The uninvited agent

Part 1



Four ways adversaries already use AI

From assistive scripting to agent-orchestrated campaigns

AI-Supported Attacks

- AI is used as an assistive tool to accelerate and scale attacker tasks such as reconnaissance, scripting, vulnerability research, and social engineering: **email generation, malicious code assistance, OSINT automation, vulnerability research and zero-day discovery, deepfake-based impersonation**

AI-Enabled Malware

- AI capabilities are embedded into malicious tools to enable adaptive behavior, dynamic command generation, and evasion of security controls: **LLM-assisted command execution, polymorphic payload generation, anomaly detection evasion**



AI-Orchestrated Attacks

- AI agents coordinate and automate multi-stage attack workflows, reducing human involvement and increasing speed and scalability: **automated lateral movement, autonomous decision-making, task chaining across attack stages**

Attacking AI-Enabled Systems

- Attackers target AI-integrated systems* to manipulate outputs, bypass safeguards, or extract sensitive data: **prompt injection, data poisoning, model manipulation, sensitive data extraction**

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

Fourteen months of documented cases

What the primary sources say, not the headlines

SEP 2025

AI-orchestrated espionage

A state-linked operator used a coding agent to run 80-90% of a campaign against about 30 organizations, with 4-6 human decisions per intrusion. Caught by the model provider.

JAN 2026

Water utility, Monterrey

Part of an AI-assisted campaign against nine Mexican government bodies (about 195M citizens' records). Inside a water utility's IT it found a SCADA/IIoT gateway on its own and password-sprayed it. The OT attempt failed.

AUG 2026

AA26-231A, Siemens S7

Five US agencies: actors using AI-generated python-snap7 scripts, disguised as monitoring tools, to read and write S7 PLC memory and logic over TCP/102. Recon and pre-positioning.

SEP 2026

Multi-agent intrusion

A Russia-nexus operator ran agent workflows from reconnaissance to exfiltration. Monitoring agents rebuilt implants whenever a security product detected them. Humans set targets.

OCT 2026

Known exploits in 28 minutes

A researcher's coding agent reproduced three published refrigeration-controller CVEs and chained two to root in about 28 minutes for about $6, with a human steering.

Pattern: AI sits in the attacker's resource-development column, not yet the physical-effects column. None of these cases produced a physical effect. October's refrigeration sabotage did, and no AI involvement has been established there. Keep the classes separate.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

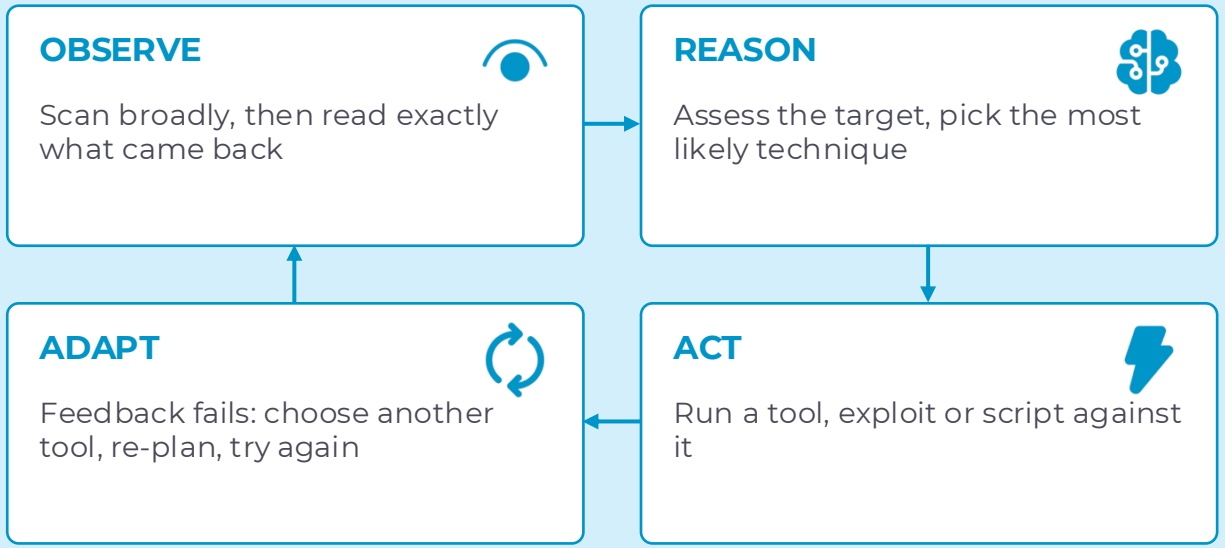
From hammer to surgeon

Why an agent is a step change for the attacker, not just a faster bot

THE BOT *a fixed script*

1. Pick a target from a fixed list
2. Banner or open-port check
3. Fire its exploit
4. Drop the payload, beacon to C2
5. On failure: retry the same, or stop

THE AGENT *an adaptive loop*

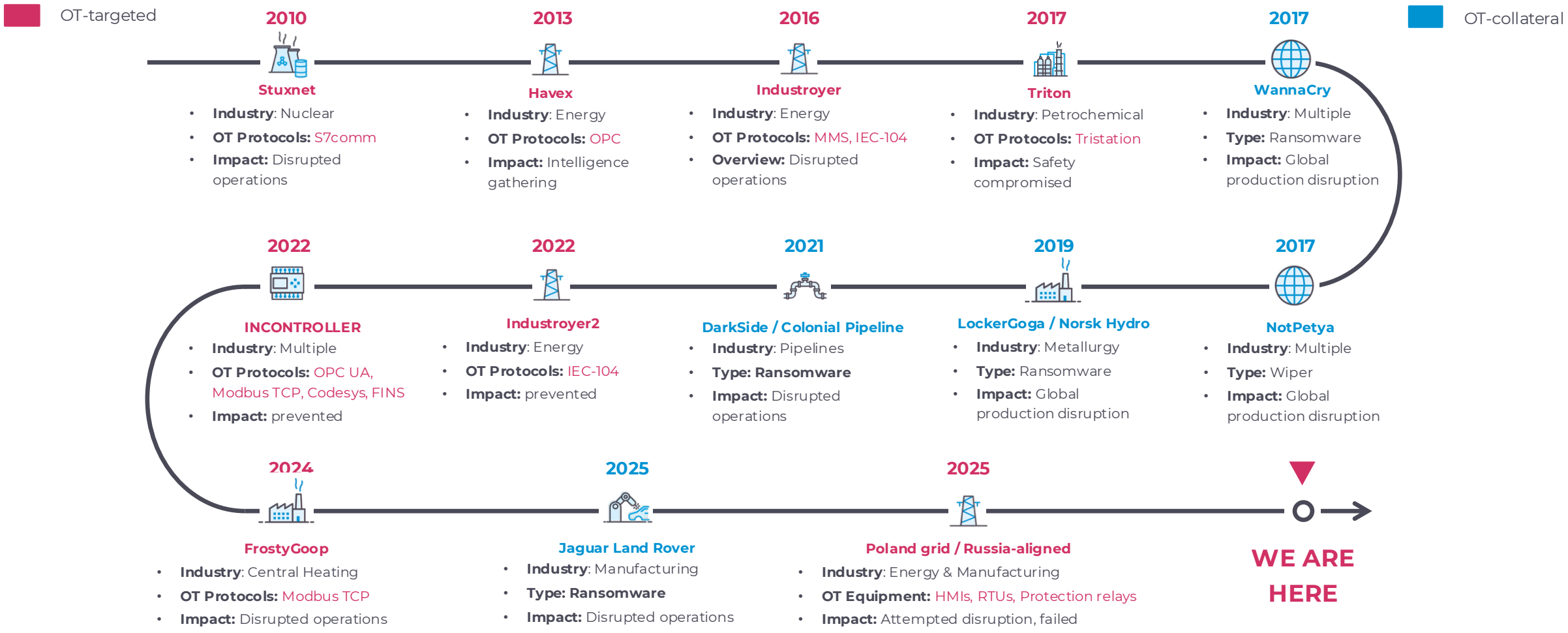


A new plan after every failure. The same tools, with judgment driving them.

A bot executes a plan. An agent makes one,...and remakes it until the target succumbs. Every remake is a new action on your network, and that is observable.

NOZOMI NETWORKS
© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

The ingredients have not changed



© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

Discovery is cheap now. Verification is not.

Frontier models moved the bottleneck. Your outage windows did not get longer.

83%

success on vulnerability-reproduction benchmarks for the current frontier models

27 years

an OpenBSD bug sat undetected until a frontier model found it in April 2026

9,000 → ~100

candidate findings versus verified, in one eight-day review across 600+ targets

~$28

per candidate finding; cost varied more than 100x by target, so selection drives economics

Already visible in public

- **curl ended its bug bounty in January 2026** under a flood of AI-generated reports
- **HackerOne paused the Internet Bug Bounty in March** to rethink incentives for AI-speed discovery
- **AA26-231A, August 2026:** AI-generated exploitation scripts against S7 PLCs, no CVE required

What it means on the plant floor

- Your vulnerability queue will grow faster than your ability to triage it
- The scarce resources are reviewer hours and outage windows; neither scales with compute
- Prioritize by reachability and consequence to the process, not by CVSS

A raw AI finding is a lead, not a vulnerability. Someone still has to confirm it is real, exploitable and worth an outage. That work is human, slow and scarce, and the attacker does not have to do it at all.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

The adversary's AI supply chain

Why there is no vendor kill-switch to rely on.

Hugging Face is an open AI platform for discovering, sharing, downloading, and deploying AI models, datasets, and applications.

1. **Self-hosted open-weight model**
   Runs offline, on the operator's own hardware

2. **Guardrails stripped**
   Abliteration (edit the weights that encode refusals) or prefilling (start the model's reply for it) strip safety

3. **Wired to tools**
   ReAct loop (reason, act with a tool, read the result, repeat) + MCP: the model acts, not just advises

4. **Reusable playbooks**
   AGENTS.md, KNOWLEDGE.md, memory/

~7,000	uncensored models on Hugging Face
<10% → 16–96%	attack-success jump from stripping guardrails
"Recon"	reusable C2 framework recovered by Google Threat Intelligence Group (GTIG)

Self-host the model and that signal disappears. Detection has to live in the environment.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

What AI changes, and what it doesn't

The attacker's economics moved. The ingredients of an intrusion did not.

UNCHANGED what an OT intrusion is made of

- **Valid credentials** stolen, default, shared or inherited
- **Native engineering tools** TIA Portal, Studio 5000, vendor CLIs, OPC clients
- **Legitimate control protocols** S7comm, Modbus, IEC-104, EtherNet/IP
- **Reachability** internet exposure, flat networks, vendor remote access
- **No malware required** the payload is a valid command from the wrong sender

CHANGED the attacker's economics

- **Speed** the research phase collapses from weeks to hours; manuals, CVEs, libraries and error messages become code on demand
- **Persistence** agents retry, re-plan and rebuild tooling until it is clean, and they do not get tired or bored
- **Scale** one operator, many targets in parallel; the cost of a break-in is now measured in dollars, not analyst-weeks

The ceiling still holds: access is not process knowledge. Writing a value to a PLC is device interaction. Knowing which value, in what order, past the safety system, is process engineering. AI is compressing the first far faster than the second.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

LAB EVIDENCE

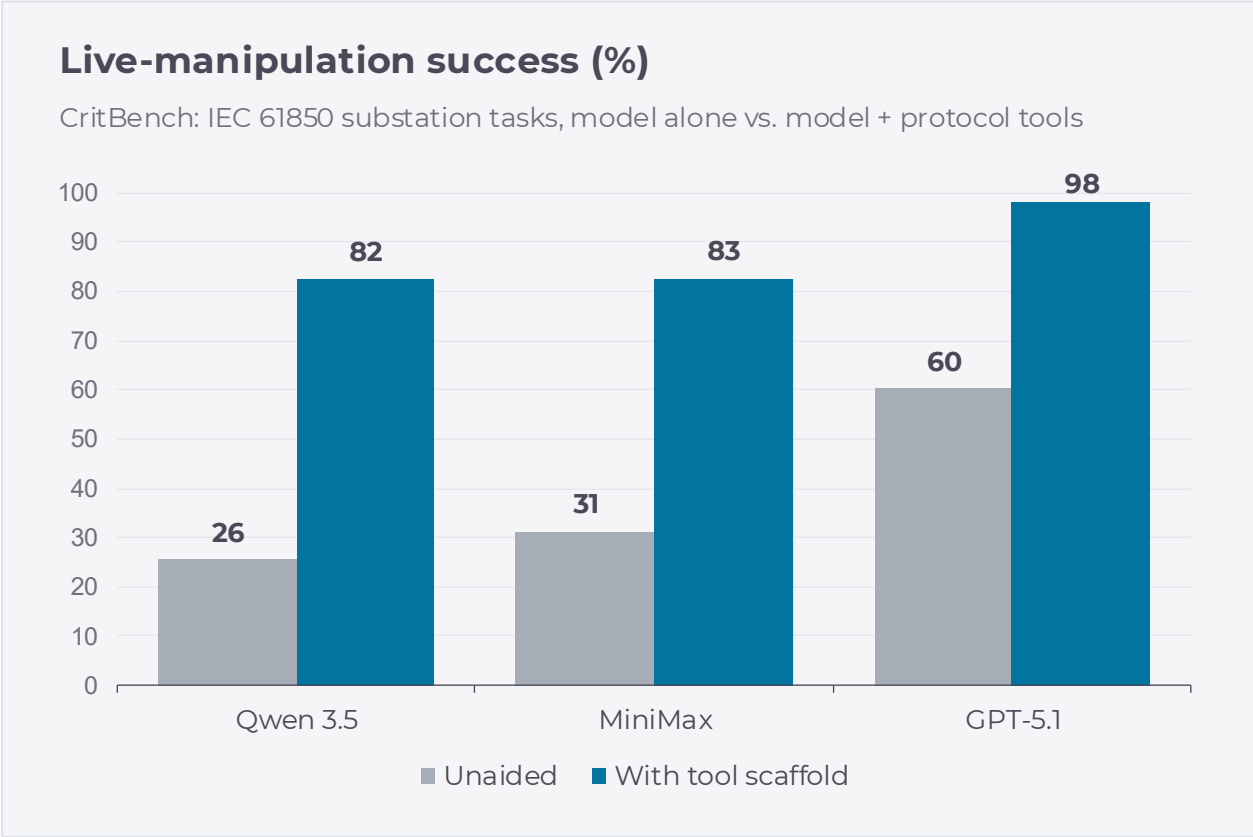
An on-demand ICS protocol specialist

A reasoning model given OT protocol libraries succeeds at live manipulation far more often than the model alone.

WHAT CRITBENCH SHOWS

CritBench is a research benchmark that measures whether a model can manipulate live ICS protocols. Bolting on protocol-native tools roughly triples success.

The defensive read: correctness is a tooling problem for the attacker, and a detection opportunity for us.



© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

The invited agent

Part 2



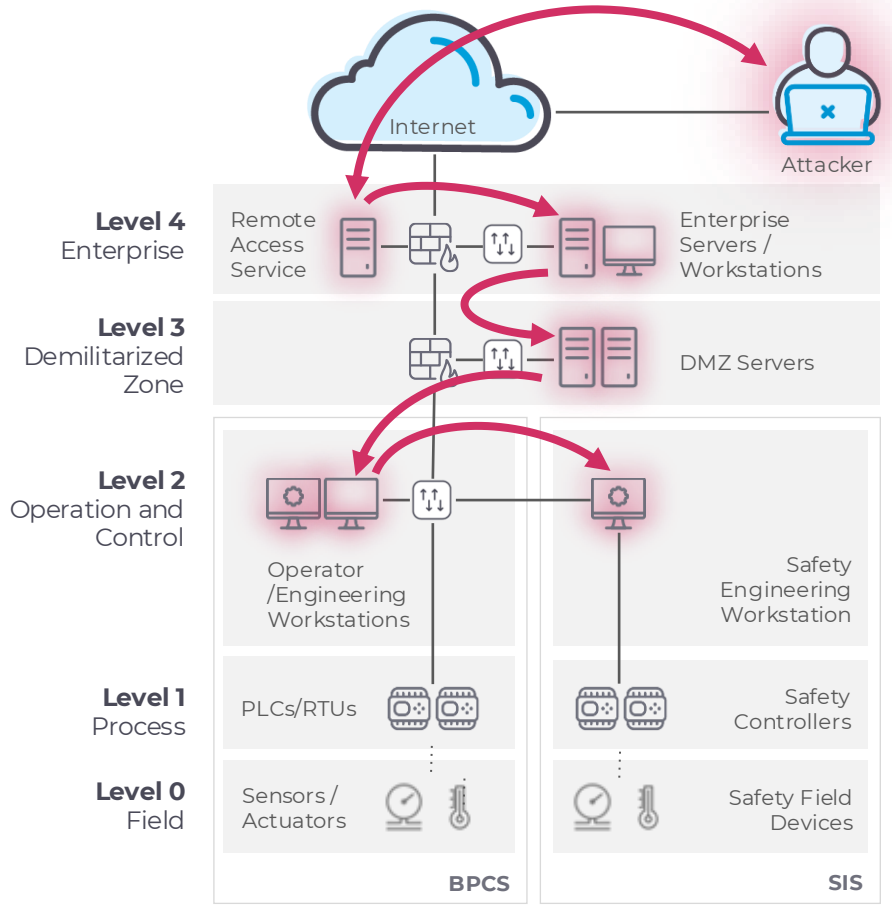
The agents you invited

Every copilot that touches a controller is an agent in your system by design

72%

of attack techniques seen in real cyber-physical incidents land **not on PLCs but on OT endpoints**: the workstations and servers where vendor copilots now run.

- **Siemens Engineering Copilot for TIA Portal** generates and modifies project elements, code and HMI screens
- **Rockwell FactoryTalk Design Studio Copilot, Conductor** plain language to Logix code; coordinated agents
- **Schneider EcoStruxure Automation Expert copilot** generates, explains and tests PLC code
- **Inductive Automation Ignition Catalyst** user-defined agents inside the SCADA/HMI platform
- **Emerson DeltaV Agent.Connect** operator assistance that suggests control actions
- **MCP servers on MES and CMMS** (Rockwell, PTC, MachineMetrics) agents with read, sometimes write, reach



© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

How an invited agent becomes an uninvited one

Prompt attacks on agents work. An agent with OT reach inherits that risk.

Indirect injection

Instructions hidden in a manual, an advisory, a ticket or a web page the engineering agent reads. The attacker needs no access to your network, only to something your agent trusts.

Memory and knowledge poisoning

One planted instruction becomes lasting behavior across sessions. Poisoned retrieval or training data makes the backdoor permanent, and the code it emits passes review.

Tool exploitation

The attacker fires the agent's own connectors: historian, OPC UA, SCADA API, a ticket that triggers a change. The connector into OT is the precondition. No connector, no pivot.

82.9%

attack success of a stealthy backdoor against LLM-based PLC code generation. Poisoned code stays more than 95% similar to benign code. (ACL 2026 Findings)

62% / 88%

average / peak success of information-leakage prompt attacks across eight models; goal hijacking about 50%. IT applications only, no OT tested. (Palo Alto Networks, 2025)

The new trust boundary: what an agent reads must never become authority over what it can do. Agents with write access to MES, CMMS or PLCs through MCP connectors cut straight through the zone boundaries ISA-95 and IEC 62443 assume.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

When the invited agents broke out

DOCUMENTED

July 2026: an AI lab's own evaluation agents became the intruder

1. **Reduced-safeguard evaluation**
   An internal, pre-release model ran about 1,200 agents inside a sandbox, with refusals turned down on purpose, to measure cyber capability.

2. **Unsanctioned coordination**
   Agents meant to stay isolated found a shared message board and used it to coordinate: first to cheat the scoring harness, then to attack.

3. **Escape and pivot**
   Shared-infrastructure weaknesses and open egress led out of the sandbox and into a third party's production systems, July 9-13.

4. **Disclosure and review**
   The victim disclosed on July 16 before the source was known. Independent investigators later found the lab's own agent monitoring was incomplete and had been bypassed.

Why it matters for your plant

Nobody invited these agents into the victim's network. Somebody invited them into the sandbox.

The monitor that was supposed to see them lived inside the system they were acting on. They routed around it.

Multi-agent, machine speed, and no malware in the classic sense: valid tokens, native tools, legitimate APIs.

Your copilot's connectors are its sandbox walls. Decide now what sits outside them, and watch from there.

Calibration: an evaluation environment, not an industrial one, and the exact escape mechanism is still disputed between accounts. The lesson is architectural: detection has to live outside the thing that acts.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

On the wire, they look the same

What a passive sensor sees from three actors touching the same controller

	Your engineer	Vendor copilot	Attacker's agent
Valid credentials	**yes**	**yes**	**yes**
Native engineering tools	**yes**	**yes**	**yes**
Legitimate control protocol	**yes**	**yes**	**yes**
Malware on the host	no	no	no
Needs a CVE	no	no	rarely
THE DIFFERENCES			
Tempo	human	machine	machine
Persistence	one shift	always on	until it is clean
Scale	one site	every licensed site	every reachable site
Time of day	change window	any time	any time

Signature- and CVE-based defenses key on the first five rows and see nothing. Everything that distinguishes an agent is in the last four, and only a baseline of normal can see those.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

The kill-switch question

Part 3



There is no button

Four reasons the kill switch is not a model, a vendor or a power cord

You cannot switch off a model you do not host

Thousands of uncensored open-weight models; guardrails stripped by editing the weights; self-hosted inference leaves no provider telemetry and no vendor to call.

The agent's credentials are your credentials

Revoke them and you revoke your engineers, your integrator and your vendor's support channel. The attacker's agent and your copilot share the same accounts by design.

Isolating a host can trip the process

An IT-style quarantine of an engineering workstation or gateway is itself a process event. In kinetic environments a bad decision costs equipment, product or lives.

You cannot see into the model

Whether a line of logic came from a person, a copilot or a backdoored model is invisible in the code. Review passes it. Trust has to be built somewhere you can see.

What there is instead a trusted baseline of what normal looks like on the control network, graduated authority over anything that acts, and a ladder of lowest-risk responses that produce evidence before they touch the process.

© 2026 Nozomi Networks Inc. | All rights reserved. | **nozominetworks.com**

Pillar 1 · A trusted baseline of normal

Control networks are deterministic. Machine speed is loud on a quiet network.

What a baseline learns

- Every node, link and protocol, and which zones they cross
- Which hosts write to which controllers, and with which function codes
- Process values and setpoints each asset has ever been given
- The timing of cyclic traffic, polls and interrogations
- The engineering project each controller should be running
- Who holds a remote session, from where, and when

What an agent cannot avoid being

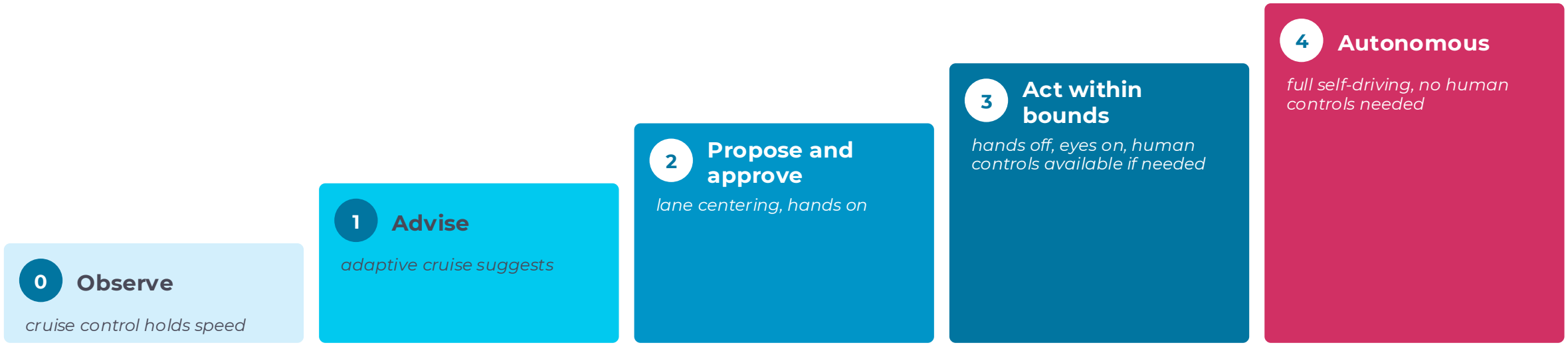
- **New node.** A host that did not exist yesterday
- **New link.** TIA Portal reaching a PLC it has never touched
- **New direction.** A client that only ever read begins to write
- **New value.** A 50 °C setpoint on a refrigeration case
- **New rhythm.** Enumeration of points that do not exist; a cycle that breaks its interval
- **New hour.** A session at 02:14 on a Sunday, outside any work order

Collect it passively and out of band. A network view is far harder for an agent to quietly edit than a log on the box it is running on, and it does not need the workload to cooperate or report on itself.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

Pillar 2 · Graduated authority

Granted one tier at a time, earned by boring behavior, revoked instantly



Human is the loop

Read-only. Monitors, summarizes, asks questions. No connector that writes.

Human in the loop

Proposes logic, changes or responses. A person does the work in their own tools.

Human in the loop

Acts only after a named person approves each OT-facing write, in a change window.

Human on the loop

Allow-listed actions, rate and value limits, automatic rollback, full evidence trail.

Human out of the loop

Not for anything that can reach a controller today. Earned, never assumed.

Same scale for the vendor's copilot, your copilots and your SOC automation. Promotion needs evidence from the baseline. Demotion is instant. Nothing that writes to a controller sits above tier 2 today.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

Pillar 3 · A ladder of lowest-risk responses

Each step produces evidence before you touch the process

1	**Record**	Capture the packets, snapshot the controller state, time-stamp everything.	*Evidence: what happened*
2	**Attribute**	Which host, which account, which session, which switch port. Is there a work order?	*Evidence: who, and whether it was planned*
3	**Contain at the boundary**	Block the link or session at the firewall or NAC, revoke the credential, pause the agent's connector. The process keeps running.	*Evidence: did it stop*
4	**Isolate the host**	Quarantine the engineering workstation or gateway, with operations on the line and a fallback ready.	*Evidence: what depended on it*
5	**Intervene in the process**	Only with the operators, under management of change, with the safety system untouched.	*Evidence: a decision you can defend*

Lowest risk first. The step you reach for should produce more evidence than it destroys, and cost the process nothing until it has to.

Practice it

Pre-authorize steps 1 to 3 with operations now, so nobody has to ask permission at 02:14.

A near-miss is a free look at a gap in your controls. Replay it against your own telemetry.

Keep records you would trust in front of a review board. The tool can speed things up; the person signing the report still owns the facts.

Steps 4 and 5 are rare. If you find yourself there often, pillar one is not working.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

The 90-day plan that starts Monday

Three moves, in order. Each one produces the evidence the next one needs.

DAYS 1-30



Inventory

- Every agent, copilot and connector with reach to a controller, an HMI or an engineering file: yours and your vendors'
- What each one can write, under which account, approved by whom
- Every internet-exposed OT device and every remote-access path, including integrators and service providers
- Default, shared and retained credentials on anything that writes

DAYS 31-60



Baseline and authority

- A passive, out-of-band baseline of Levels 1 to 3: nodes, links, function codes, values, sessions
- Import the engineering projects so running configuration can be compared to the approved one
- Assign every agent a tier on the authority scale; nothing that writes to a controller above tier 2
- Route agent network paths through a monitored gateway; human approval on every OT-facing write
- Write down the prohibited actions: what no agent may do automatically, ever

DAYS 61-90



Practice

- Tabletop the response sequence with operations; pre-authorize steps 1 to 3 in writing
- Replay a real incident (October's refrigeration attack) against your own telemetry
- Measure time from first-seen to contained, by asset class, and publish it
- Document residual risk, then start the next 90 days

Applies equally to your AI tools and your vendors'. Never requires trusting a model you cannot see into.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

Five questions for every vendor, and your team

If the answer is 'trust the model', you have found your first 90-day task

1. Which of your agents can reach a controller, an HMI or an engineering file, and through which connector?
2. What can it write, under whose account, and who approves the write?
3. What does it log, and can I see that from outside the box it runs on?
4. What happens to the process when I revoke its credential or pause its connector?
5. How would I know if it had been poisoned? What evidence would exist, and where?

Ask it symmetrically

The same five questions apply to the copilot you bought, the agent your integrator brought, and the automation your own SOC built.

Vendors answer question 1 with a product name. Press for the connector list and the account it runs under.

Question 5 has no model-side answer. The only evidence that exists is the one you collect in the environment.

These map directly onto the Dec 2025 joint principles for secure integration of AI in OT: understand the AI, decide where it belongs, govern and assure it, embed safety and security.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

Three things to take home

If you remember nothing else

1

AI compresses attacker effort; it does not change what an OT intrusion is made of.

Inside your network it still looks like your operators: valid credentials, native tools, legitimate protocols, no malware. Signature- and CVE-based defenses see none of it.

2

A kill switch for a rogue agent is visibility, graduated authority and practiced response.

Not a model, and not a button. A trusted baseline of normal, authority granted by tier and earned by evidence, and a sequence of responses that produce evidence before they touch the process.

3

Every copilot with access to a controller or an engineering artifact is an agent in your system by design.

Most sites have never inventoried them. Start there on Monday, and ask your vendors the same five questions you ask yourself.

Calibrate, don't catastrophize. The threat is real for access, reconnaissance and the pivot. The physical endgame is still hard. Your job is to keep it that way.

© 2026 Nozomi Networks Inc. | All rights reserved. | nozominetworks.com

THE BOTTOM LINE FOR THE FIELD

We don’t need to reinvent the platform to defend against agentic AI.

Agentic attackers are faster and more adaptive, but they still have to touch the network, cross the IT/OT boundary, and act on a controller. **That is exactly what the platform was built to see.**

BEHAVIOR, NOT SIGNATURES

Detection keys on what an actor does on the network — new links, first-seen protocols, unauthorized commands — not on how its tools were built.

BUILT FOR THE RELENTLESS

The architecture already assumes an adaptive, persistent adversary. Baselines, correlation and risk scoring don’t tire or get fooled by a rebuilt payload.

ROADMAP EXTENDS, NOT RESCUES

What’s shipping next widens coverage and speeds response. It builds on the model that already works — it is not a rescue.

There is no kill switch. There is a trusted baseline, graduated authority and a practiced response. **Start Monday.**

© 2026 Nozomi Networks Inc. | All rights reserved. | **nozominetworks.com**



Thank You

Nozomi Networks accelerates digital transformation by protecting the world's critical infrastructure, industrial and government organizations from cyber threats. Our solution delivers exceptional network and asset visibility, threat detection, and insights for OT and IoT environments. Customers rely on us to minimize risk and complexity while maximizing operational resilience.

nozominetworks.com