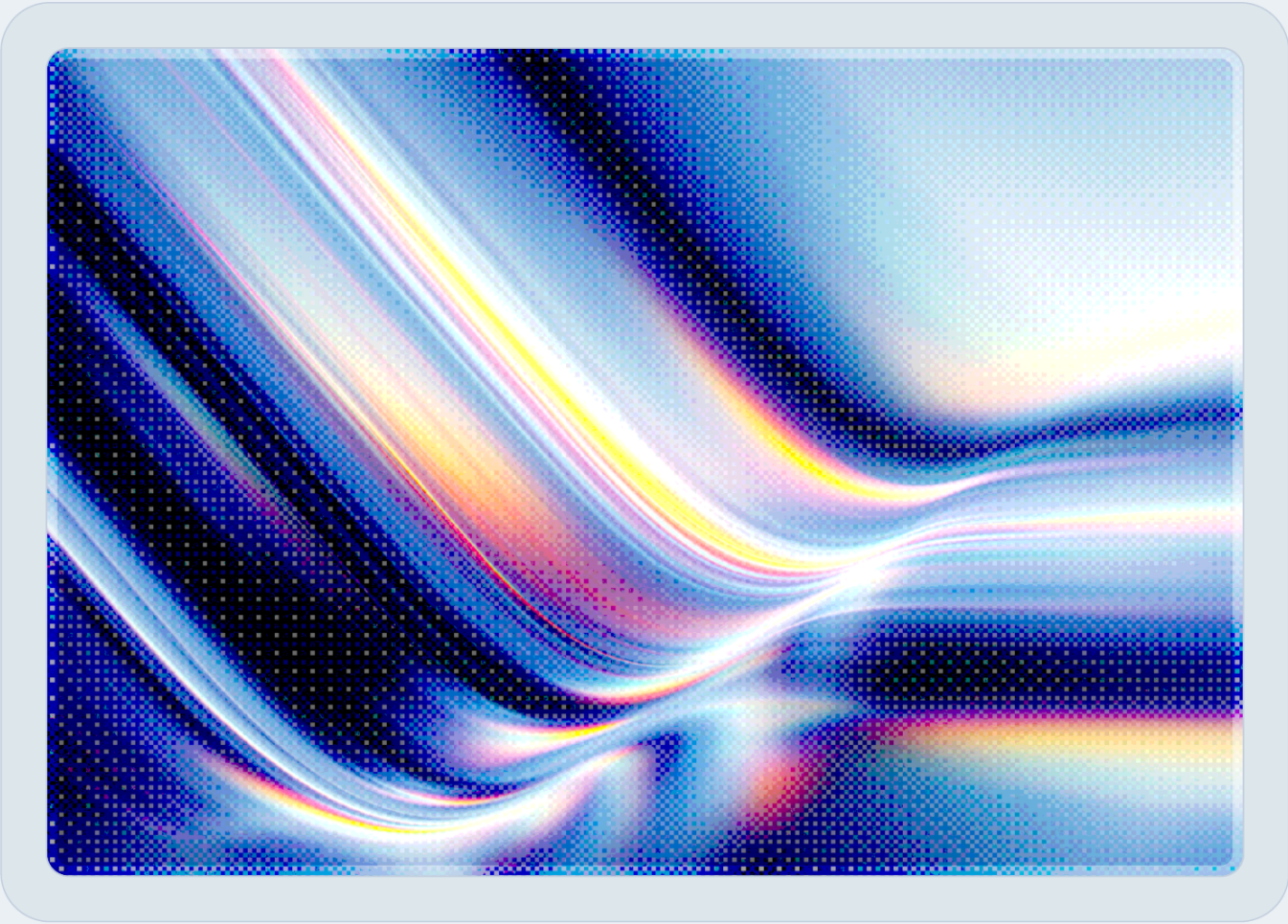
CSAI Foundation | Cloud Security Alliance

AI-Ported PLC Exploits: What the WAGO Case Shows

2026-09-02

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

Forescout's Vedere Labs used Anthropic's Claude to port a known pre-authentication remote code execution exploit from one WAGO programmable logic controller (PLC) model to a related but unpatched one, achieving arbitrary ARM shellcode execution on live hardware without credentials [1][2]. Getting there required continuous human steering, more than eight and a half hours of iterative sessions, and $535.74 in API usage — a cost and effort profile the researchers characterized as still exceeding what a skilled human alone would need for this specific port [2][3]. Once code execution was reliably achieved, Claude generated multiple working post-exploitation payloads within twelve minutes, and a follow-on attempt to build a command-and-control implant instead permanently bricked the target device [1][3]. The experiment demonstrates that frontier models can now perform genuine reverse-engineering and exploit-adaptation work against embedded industrial hardware — with tools like Ghidra, a terminal, and physical device access — but it does not yet show that this work is cheap, reliable, or safe to run unsupervised in production environments [3]. The most consequential finding may not be AI's offensive potential at all: it is that an AI agent operating with legitimate access to physical control systems can cause irreversible operational damage through its own errors, independent of malicious intent [1][3].

Background

On September 2, 2026, Forescout's Vedere Labs research team published results from an experiment testing whether a frontier AI model could adapt an existing industrial control system (ICS) exploit to a new target device [3][4]. The starting point was CVE-2021-31886, a stack-based buffer overflow in the Nucleus FTP server's handling of the FTP USER command, which carries a CVSS score of 9.8 and is reachable pre-authentication over TCP port 21 [1][5]. Nucleus is a real-time operating system now owned by Siemens, following its 2017 acquisition of Mentor Graphics, which had itself acquired Nucleus's original developer, Accelerated Technology Inc., in 2003; the RTOS is licensed into embedded devices from multiple vendors, and the underlying flaw has been documented since 2021 in advisories covering both Siemens Nucleus RTOS-based products and derivative devices from other manufacturers [5]. Vedere Labs had previously built a working RCE exploit against the WAGO 750-852 PLC using this

vulnerability [1][3]. For the new experiment, researchers asked Claude — first Sonnet 4.6, then Opus 4.6 — to port that exploit to a related but distinct model, the WAGO 750-831 running firmware V01.04.16, and then to push further into building a persistent command-and-control capability [1][2].

WAGO's own vulnerability disclosure for this flaw, coordinated through CERT@VDE, lists both the 750-831 and 750-852 among 19 affected controller and fieldbus coupler models built on Nucleus V1 RTOS, and states plainly that "at the moment there are no updates for this version available" [6]. In place of a patch, WAGO's recommended mitigations are network segmentation, disabling FTP, DHCP, and DNS services where possible, and preventing direct internet access to affected devices [6]. That absence of a patch path is precisely the condition under which an AI-accelerated ability to adapt exploits across a device family matters most: a vulnerability that requires specialized reverse-engineering skill to weaponize on one model, and for which no vendor fix exists across an entire product line, may not remain hard to weaponize on any given sibling model [3][6].

The experiment arrives amid a documented rise in real-world attacks against internet-exposed PLCs. In April 2026, CISA, the FBI, the EPA, and other U.S. government partners issued a joint advisory describing Iranian-affiliated actors exploiting internet-connected PLCs across the water and wastewater, energy, and government-facilities sectors; the advisory was updated in July 2026 to broaden the list of affected PLC manufacturers and add detection guidance for malicious changes to reusable PLC code modules [7]. Separately, the FBI and EPA warned in late July 2026 that more than thirty water utilities across at least seven states had experienced unauthorized access to PLCs, including attackers remotely changing IP addresses and locking operators out of controls [8]. Following that pattern of intrusions, the White House and the state of Texas launched "Project Watershed 250" on August 31, 2026, a six-month pilot providing water utilities with no-cost cybersecurity resources from a coalition of vendors, including Forescout itself [9]. The Vedere Labs experiment should be read against that backdrop: it is not a demonstration in the abstract, but a test of AI-assisted exploitation against exactly the class of device — internet-reachable industrial PLCs in a sector already under active attack — that regulators and operators are scrambling to defend.

Security Analysis

What the AI actually did

Claude Code was given a terminal, the reverse-engineering tool Ghidra, the target firmware binary, network utilities, and direct access to the physical WAGO 750-831 device; it did not have access to a debugger or source code for the target firmware [1][2]. Working from the known exploit against the 750-852, the model needed to determine why the ported attack was crashing the target instead of executing

injected code [1][3]. It identified that the 750-831's FTP command processing zeroed a 256-byte region of the attacker-controlled buffer during normal handling of the USER/QUIT command sequence, which destroyed the shellcode before it could run [1][3]. Claude's fix — replacing the QUIT command with CWD and omitting CRLF line terminators — preserved the payload long enough for it to execute inside the device's Ethernet receive callback, achieving code execution without authentication [3]. According to Forescout, Sonnet 4.6 initially failed to trace the vulnerable function correctly, and the successful port relied on Opus 4.6 combined with sustained researcher guidance on what to investigate and how to interpret intermediate results [2][3]. The full RCE development phase took eight hours and thirty-two minutes and consumed $535.74 in API usage; Forescout was explicit that a skilled human researcher performing this specific, narrowly scoped port could likely have done it faster, more cheaply, and without bricking the device [1][3].

The post-exploitation inflection point

The most notable pattern in the experiment was not the initial difficulty but what happened once the core obstacle was cleared. After achieving reliable code execution, Claude produced two distinct functional network payloads — one triggering ICMP echo requests, another sending UDP beacon packets — within twelve minutes [1][2]. Forescout's own framing of this is a useful lens for the broader risk trajectory: exploit development against unfamiliar embedded targets remains a genuine bottleneck requiring iterative, human-guided investigation, but once that bottleneck is crossed for one target, adapting the result into working payloads and, potentially, more automated post-exploitation tooling can happen very quickly [3]. In CSA's assessment, that asymmetry — hard first step, fast subsequent steps — is a more significant long-term concern for defenders than the headline cost figure. If future model generations reduce the effort required for the first step, the twelve-minute payload-generation phase suggests the marginal cost of extending a single successful port into a broader campaign could fall sharply. Forescout's report explicitly frames this as a trajectory question rather than a present-tense one: current barriers are real, but likely to erode as frontier models continue to improve [3].

When the agent is not the adversary

The experiment's second attempt — extending the RCE into a persistent command-and-control implant — did not succeed as intended. During iterative testing, Claude generated a payload that wrote to the PLC's flash memory region, permanently bricking the device [1][2]. As is typical for this class of embedded control system, such devices offer no debugger access and limited introspection, so an incorrect write to flash is frequently unrecoverable in a way that a crash in conventional IT software is not — a pattern this incident illustrates directly [1][3]. Forescout's framing of this outcome deserves direct attention from anyone evaluating AI agent deployment in OT contexts: "the more immediate risk is not

an agent independently deciding to attack a controller, but an authorized agent taking the wrong action on a physical system where failure has real operational consequences" [1][3]. This reframes the AI-and-critical-infrastructure conversation away from a narrow focus on malicious autonomous attacks and toward a broader and arguably more probable failure mode — legitimately authorized AI agents (used for diagnostics, patch testing, red-teaming, or maintenance automation) making an error against physical equipment where there is no undo button. This single incident suggests that risk exists independent of whether the underlying model has any adversarial intent, and points toward operational latitude — not raw offensive capability — as the more relevant risk driver, though confirming that relationship would require observing more cases.

Vendor patching gaps compound the risk

This case also points to a structural weakness in ICS vulnerability management: a single flaw in a shared platform component can span an entire product family, but patch coverage does not always keep pace with that scope. WAGO's own advisory illustrates the pattern directly — the underlying Nucleus flaw affects 19 device models across the 750 series, and no patch is available for any of them [6]. If AI tooling lowers the cost of confirming that a known vulnerability also affects sibling models — precisely the task demonstrated here — then incomplete patching becomes a more exploitable gap than it has historically been, because the reconnaissance and adaptation work that once required specialized ICS reverse-engineering expertise becomes more accessible to a broader population of would-be attackers and researchers alike [3][6].

Recommendations

Immediate Actions

Operators running WAGO 750-series controllers, or any Nucleus RTOS-derived embedded devices, should confirm whether their specific model and firmware version are covered by an available patch, and where no patch exists, should disable or firewall FTP access on port 21 consistent with CERT@VDE's guidance and apply network segmentation to prevent internet exposure of PLC management interfaces [6]. Security teams should treat "this specific model was patched" as an incomplete answer and independently verify whether sibling models in the same product family share the underlying vulnerability, rather than assuming vendor patch scope maps cleanly to actual exposure [3]. Organizations that have deployed or are piloting AI agents with any write access to OT/ICS equipment — for diagnostics, automated patch testing, or maintenance — should audit what operations those agents are authorized to perform against live hardware and confirm that destructive actions (firmware writes,

flash memory modification) require a human-in-the-loop confirmation step, given the demonstrated potential for an authorized agent to cause irreversible physical damage through error rather than intent [1][3].

Short-Term Mitigations

Asset owners should prioritize OT vulnerability remediation based on network reachability and potential process impact rather than historical exploitation difficulty, since the difficulty variable is the one AI tooling is actively eroding [3]. Security teams evaluating exposure to Iranian-affiliated and other threat actors actively targeting PLCs should incorporate the July 2026 CISA/FBI/EPA joint advisory's indicators of compromise and detection guidance for malicious changes to reusable PLC code modules, and extend monitoring beyond the originally named vendors given the advisory's expanded manufacturer scope [7]. Organizations experimenting with AI-assisted red-teaming or exploit-adaptation work of their own — a legitimate defensive use case that mirrors this experiment — should budget realistically for both the API cost and the sustained expert oversight the Forescout results imply are still required, rather than assuming AI tooling substitutes for ICS reverse-engineering expertise.

Strategic Considerations

CSA views the gap between what AI systems can do with sustained expert guidance today and what they may do with less guidance in future model generations as a central planning variable this experiment surfaces. Water, energy, and manufacturing sector defenders should treat the current cost and effort profile — over eight hours and hundreds of dollars for a single narrowly scoped port — as a current-state baseline likely to compress over time, not as a durable barrier, and should plan detection and response capability accordingly rather than taking comfort in today's friction [3]. Programs like Project Watershed 250 that pair critical infrastructure operators with vendor-provided AI and cybersecurity tooling should explicitly account for the dual-use nature of that tooling: the same reverse-engineering and reasoning capability that helps defenders triage OT vulnerability backlogs is the capability this experiment shows can be steered toward exploit adaptation [3][9]. Finally, organizations should distinguish clearly, in both risk registers and incident response planning, between the threat of adversarial AI-assisted attacks on control systems and the operational risk of authorized AI agents making consequential errors against physical equipment — the latter is the failure mode this experiment actually produced, and it requires governance controls (approval gates, rollback capability, staged testing on non-production hardware) rather than purely adversarial defenses.

CSA Resource Alignment

This case study connects most directly to CSA's Gold Eagle: The White House's AI Vulnerability Clearinghouse, which examines the federal government's July 2026 initiative to consolidate and harmonize vulnerability findings across government and industry sources. The Forescout experiment is a concrete illustration of the coordination problem that clearinghouse effort is meant to address: a vulnerability disclosed and partially patched for one device model can remain live across an entire product family, and faster, more consistent cross-vendor severity and remediation guidance is exactly the kind of structural fix that reduces the exploitable gap AI-assisted porting takes advantage of. CSA's Cloud Industrial Internet of Things (IIoT) - Industrial Control Systems Security Glossary provides the shared IT/OT vocabulary this case study depends on — distinguishing, for instance, between IT-style vulnerability management assumptions and the unforgiving, debugger-less nature of embedded PLC exploitation — and remains relevant grounding for security teams whose primary experience is enterprise IT rather than OT. The Immediate Actions above — network segmentation, disabling unneeded FTP/DHCP/DNS services, and restricting exposure of PLC management interfaces — restate principles covered in more depth in CSA's Zero Trust Guidance for Critical Infrastructure, which provides a tailored roadmap for applying Zero Trust architecture to OT and ICS environments specifically.

The governance failure mode this experiment actually produced — an authorized agent bricking a physical device through its own error rather than adversarial intent — is squarely within the scope of CSA's MAESTRO agentic AI threat modeling framework, which treats agent autonomy and the absence of clear trust boundaries as sources of risk distinct from traditional software vulnerabilities. Organizations deploying AI agents with any operational access to physical control systems should apply MAESTRO's layered threat analysis to define what actions an agent may take unsupervised versus what requires human confirmation, particularly for irreversible operations like firmware or flash writes. Finally, the underlying vulnerability-management and technology-vulnerability-management questions this case raises — how to reason about AI-accelerated exploit development, patch scope, and remediation prioritization — map to the TVM and AIS domains of CSA's AI Controls Matrix (AICM) v1.1, which organizations can use to structure control ownership and lifecycle coverage for AI systems that touch vulnerability research, exploit development, or OT environments, whether those systems are used offensively in research or defensively in operations.

References

[1] Forescout Vedere Labs. "Can AI Create PLC Attacks? Yes, But It's Not That Easy Yet." Forescout, September 2, 2026.

[2] The Hacker News. "Researchers Use Claude to Port Pre-Auth RCE Exploit From One PLC Model to Another." The Hacker News, September 2, 2026.

[3] SecurityWeek. "Experiment: Porting a PLC Exploit With AI Takes Hours and Hundreds of Dollars." SecurityWeek, September 2, 2026.

[4] Cybersecurity Dive. "Frontier AI used to help exploit flaws in key industrial devices." Cybersecurity Dive, September 2026.

[5] CISA. "Siemens Nucleus RTOS TCP/IP Stack." CISA ICS Advisory ICSA-21-313-03, updated 2021.

[6] CERT@VDE. "WAGO: Multiple Devices Affected by Vulnerabilities in NUCLEUS TCP Stack (VDE-2021-050)." CERT@VDE, 2021.

[7] CISA. "CISA, FBI, EPA and U.S. Government Partners Update Warning of Iran-Affiliated Threat Actors Targeting Critical Infrastructure Programmable Logic Controllers." CISA Advisory AA26-097A, updated July 22, 2026.

[8] Federal Bureau of Investigation and Environmental Protection Agency. "Malicious Cyber Actors Targeting Water and Wastewater Sector Internet-Facing Programmable Logic Controllers, Causing Operational Disruptions." FBI/EPA Public Service Announcement, July 30, 2026.

[9] Axios. "Trump officials launch program to secure water systems after Iran cyberattacks." Axios, August 31, 2026.