CSAI Foundation

Cloud Security Alliance AI Safety Initiative

ZionSiphon: AI-Assisted ICS Sabotage Targeting Water Infrastructure

Prototype Malware Embeds LLM Operator Intelligence Against Israeli Desalination and Water Treatment Systems

Unofficial AI-assisted Research

2026-04-21

Key Takeaways

- ZionSiphon is a purpose-built operational technology malware first detected on VirusTotal on June 29, 2025, days after the Twelve-Day War between Iran and Israel (June 13–24, 2025). Researchers at Darktrace analyzed the sample and identified intent to sabotage Israeli water treatment and desalination facilities — specifically the Mekorot national water carrier and major desalination plants at Sorek, Hadera, Ashdod, Palmachim, and the Shafdan wastewater treatment complex. [1][2]
- The malware implements a two-stage geographic and environmental filter before activating, then attempts protocol-specific manipulation over Modbus (port 502), DNP3 (port 20000), and S7comm (port 102). The most developed attack path targets Modbus-connected devices with instructions designed to force elevated chlorine dosing, open valves, enable pumps at maximum flow, and increase reverse osmosis pressure — modifications that, depending on plant-specific configurations and safety interlock states, could render treated water unsafe or damage physical equipment. [1][3]
- ZionSiphon's command-and-control architecture includes a Model Context Protocol (MCP) server that feeds stolen operational data to a large language model, providing the human operator with AI-generated analysis of environmental reconnaissance results and assessments of exploitable conditions. [1][4][5] Code analysis also indicates hallmarks of LLM-assisted development in portions of the malware itself, consistent with Darktrace's documented observation of AI/LLM-generated malware samples in the wild in 2026. [6]
- The current sample is non-functional due to an XOR key mismatch in its country-validation logic — the malware triggers a self-destruct routine rather than its payload — but this defect and the presence of incomplete DNP3 and S7comm implementations indicate active development rather than an abandoned project. [1][3]
- Water and wastewater utilities should treat this disclosure as a call to audit OT network segmentation and USB device policies immediately, particularly any facility whose control network has exposure to corporate IT or internet-connected jump hosts.

Background

The discovery of ZionSiphon arrived in a geopolitically charged moment. On June 13, 2025, Israel and Iran entered a brief but intense period of direct hostilities, including ballistic missile exchanges and strikes on infrastructure, that lasted until a ceasefire on June 24 — widely referred to as the Twelve-Day War. [1][2] Within days of that ceasefire, a malware sample was submitted to VirusTotal bearing markers consistent with Iranian hacktivist development and carrying encoded strings that translate to "Poisoning the population of Tel Aviv and Haifa." Additional embedded text claimed solidarity with Iran, Palestine, and Yemen. [1][2]

The targeting logic makes the intent explicit. The malware compares the infected host's IP address against a hardcoded list of Israeli network ranges — including blocks such as 2.52.0.0–2.55.255.255, 79.176.0.0–79.191.255.255, and 212.150.0.0–212.150.255.255 — before scanning for process names, directories, and file paths associated with reverse osmosis, desalination, chlorine handling, and plant control systems. [1][3] These checks implement a function Darktrace analysts named `IsTargetCountry()` in their writeup, paired with `IsDamDesalinationPlant()` to ensure the payload activates only when both conditions are simultaneously satisfied. This dual-gating approach reduces the chance of premature detonation in unintended environments and maximizes the probability that, when the payload executes, it is operating inside an active water treatment facility — a functional design choice consistent with intentional targeting discipline.

The broader threat context in which ZionSiphon appears is not new. Iran-affiliated group CyberAv3ngers, tracked by Microsoft as Storm-0784, Mandiant as UNC5691, and Dragos as Bauxite, has been conducting sustained campaigns against water and wastewater systems since at least November 2023, when it compromised Israeli-made Unitronics Vision Series PLCs at U.S. municipal water facilities. [7] By April 2026, a joint advisory from six U.S. agencies — FBI, CISA, NSA, EPA, DOE, and Cyber Command's CNMF — confirmed that the same threat actor had escalated to exploiting authentication bypass vulnerabilities in Rockwell Automation controllers and deploying the custom ICS malware platform IOCONTROL across U.S. water, wastewater, and energy infrastructure, with confirmed operational disruption and financial loss. [8]

ZionSiphon differs from IOCONTROL in scope and target geography but fits squarely within the same sustained campaign by Iranian-linked actors to translate geopolitical hostility into credible destructive capability against water-sector OT environments.

Security Analysis

Operational Technology Targeting Capabilities

ZionSiphon's OT-specific logic centers on three industrial protocols. After establishing persistence via a Windows registry run key and obtaining administrator privileges, the malware enumerates the local subnet looking for open ports associated with ICS communications. For Modbus-connected devices, the malware sends a specific register-read request (`01 03 00 00 00 0A`) and parses the response to understand register layout before issuing write commands. [1][3] The write instructions Darktrace observed are designed to inject values that would force unsafe operating states: elevated chlorine concentrations, maximum pump output, fully open valve positions, and higher reverse osmosis system pressure — parameters that if sustained would either damage mechanical systems or produce water unfit for human consumption.

The DNP3 and S7comm code paths are structurally present but functionally incomplete. The DNP3 branch returns a fixed byte sequence rather than a properly formed command frame, and the S7comm branch similarly lacks the session-establishment handshake required to communicate meaningfully with a Siemens controller. [1][3] This pattern — one mature attack path and two skeletal ones — is consistent with a phased development approach where the Modbus capability serves as a working proof-of-concept while the Siemens and DNP3 capabilities are being built out. It does not suggest the threat actor lacks the knowledge to complete them; it suggests they have not yet had time to do so.

The malware's USB propagation mechanism copies a hidden executable named `svchost.exe` onto removable media and creates Windows shortcut files configured to execute it silently when clicked. [2] [3] This propagation vector is specifically relevant to water treatment environments because many operational technology networks in the water sector maintain limited or segregated connectivity to IT infrastructure, and removable media remain an operational transfer mechanism in such environments — a documented vector in prior ICS compromise scenarios.

A significant defect in the current sample prevents execution: the `IsTargetCountry()` function applies XOR encoding to the string "Israel" and compares the result against a separately hardcoded byte sequence, but the two encodings were generated with different keys and therefore never match. [1] The function always falls through to `SelfDestruct()`, meaning the malware deletes itself rather than activating. The XOR key mismatch is consistent with an implementation error that could arise from development under time pressure or across multiple contributors. The defect is unlikely to persist in a subsequent release; analysts should treat the bug as a versioning artifact rather than a design-level failure.

The AI Layer: LLM-Augmented Command and Control

A notable distinguishing feature of ZionSiphon relative to earlier ICS malware campaigns is the architecture of its command-and-control infrastructure. The malware reports collected environmental data — process listings, file system findings, and network reconnaissance results — to an MCP (Model Context Protocol) server operated by the threat actor. An LLM analyzes this incoming data and generates a natural-language summary for the human operator, including an assessment of what has been stolen, which systems appear most exploitable, and what the operational environment looks like. [1][4][5]

This architecture shifts the attacker's operational tempo. In previous ICS campaigns, human operators needed to manually sift through reconnaissance dumps to understand what the compromised host could reach. An LLM-augmented C2 server could substantially compress the analysis step that would otherwise require manual triage, and lowers the expertise threshold required to interpret raw OT reconnaissance output — though the degree of compression will depend on model capability and deployment infrastructure. A handler who lacks deep knowledge of Modbus register semantics or desalination plant process configurations can receive a plain-language briefing from the LLM layer and make targeting decisions accordingly. The MCP protocol, originally designed to give AI agents structured access to tools and data sources, is here being repurposed as an intelligence channel between malware and operator — a convergence of AI tooling and adversarial infrastructure with no clear precedent documented in available ICS threat reporting to date.

Beyond the C2 architecture, code analysis of ZionSiphon reveals structural and stylistic markers consistent with LLM-assisted development. Darktrace has separately documented the existence of malware samples in the wild whose code bears hallmarks of AI generation — consistent comment density, standardized error-handling patterns, and modular structure that differs from traditionally handwritten malware — and has published analysis of one such sample exploiting the React2Shell vulnerability. [6] ZionSiphon exhibits similar characteristics. Whether the authors used an LLM to scaffold initial code, to implement individual protocol handlers, or to refine existing code is not confirmed from available samples, but the combination of these indicators with the operator-side LLM infrastructure suggests a development workflow in which AI tooling was present at multiple stages.

Geopolitical Context and Threat Actor Attribution

The preponderance of technical and behavioral evidence points to Iranian-aligned development, though direct attribution to a specific group has not been publicly confirmed at the time of writing. The malware's ideological content, its timing relative to the Twelve-Day War, and its operational focus on Israeli national water infrastructure collectively fit the pattern established by CyberAv3ngers and related

IRGC-affiliated actors. [7][8] Nation-state and hacktivist attacks on operational technology environments doubled in 2025 relative to 2024 according to the Waterfall Security Threat Report 2026, a vendor-published survey of publicly reported OT security incidents, with five of the fourteen documented attacks in 2025 directly linked to the kinetic conflict of the ongoing Russia-Ukraine war and a further cluster tied to Middle East hostilities. [9]

Water infrastructure has become a disproportionately targeted sector in this environment. The combination of historically accessible attack surfaces — internet-exposed OT assets have often presented low-barrier entry points — and the disproportionately severe public health consequences of successful manipulation likely makes the sector attractive for coercive signaling campaigns, though individual actor motivations vary. Disruption of water treatment threatens direct harm to civilian populations, making it useful for coercive signaling even before a payload deploys. ZionSiphon's encoded message about poisoning Tel Aviv and Haifa suggests the authors intend exactly this kind of coercive effect.

Recommendations

Immediate Actions

Organizations operating water treatment and desalination infrastructure should take several steps without delay. Any OT control system with an active USB interface should have that interface disabled or physically blocked unless it is currently required for a specific operational task. The ZionSiphon USB propagation mechanism does not require network access to the OT environment, meaning it bypasses network segmentation controls if personnel carry drives between zones.

Operators should also audit their OT networks for unauthorized listening services on ports 502 (Modbus), 20000 (DNP3), and 102 (S7comm). While these are legitimate industrial protocol ports, their exposure across inter-zone boundaries or on hosts that should not be communicating with field devices is a finding that warrants investigation. Any host in the OT environment that is making outbound connections to external IP addresses — particularly using protocols consistent with MCP or HTTP-based C2 — should be treated as a potential indicator of compromise.

IT/OT boundary controls deserve immediate review. ZionSiphon's geographic targeting logic relies on the infected system falling within Israeli IP space, which suggests initial infection occurs somewhere on the corporate network or a system with internet connectivity, not deep inside an air-gapped control network. The path from that initial foothold to OT systems is the critical junction. Organizations that have not conducted a recent network segmentation review should prioritize one.

Short-Term Mitigations

Over the next thirty to ninety days, water sector operators should implement or validate behavioral monitoring for ICS-relevant protocols on their OT networks. A Modbus message that writes to registers controlling chlorine dosing or valve positions from a host that has no business doing so is detectable — but only if there is monitoring infrastructure in place to observe it. ICS-aware network detection and response (NDR) tools should be deployed in passive monitoring mode at minimum, with alerting configured for protocol anomalies.

Incident response plans should be reviewed against the specific scenario of OT sabotage. Many water utilities have business-continuity plans that address ransomware or IT outages, but fewer have exercises that simulate physical-consequence scenarios — what the facility does if chlorine injection controls are manipulated and the water is already in the distribution network. Tabletop exercises built around the ZionSiphon sabotage scenario would surface gaps in both the technical response workflow and the public health notification chain.

Threat intelligence sharing with sector peers and CISA's Water and Wastewater Systems Sector-Specific Agency (EPA) should be activated where it has not already been. The ZionSiphon indicators of compromise from Darktrace's analysis — IP ranges, process names, file-system artifacts, protocol signatures — should be imported into any deployed threat intelligence platforms.

Strategic Considerations

The integration of LLM capabilities into malware command-and-control architecture represents a qualitative shift in the threat model for critical infrastructure operators, not merely a tactical evolution. When AI-generated intelligence reduces the expertise threshold for conducting effective OT attacks, the pool of actors capable of mounting consequential campaigns expands. ICS malware development has historically required rare expertise in both industrial protocols and software development; AI tooling erodes that barrier.

Organizations should begin evaluating AI-enabled defensive capabilities to match. Behavioral analytics platforms that establish OT traffic baselines can detect novel protocol anomalies that predefined signatures miss — particularly relevant here because ZionSiphon's register-write patterns would not match existing ICS malware signatures. Organizations evaluating such tools should weigh detection capability against false-positive rates in their specific OT environment. Monitoring for MCP-based outbound communications — a protocol with no currently established legitimate use case in standard OT environments — should be added to detection rulesets. Any MCP traffic observed in OT network zones warrants investigation.

At the strategic level, water sector operators should engage with the regulatory and information-sharing frameworks that are being updated in response to the current threat environment. CISA's Binding Operational Directive framework and cross-sector cybersecurity performance goals provide actionable benchmarks. The ISA/IEC 62443 standard series for industrial automation and control system security is a widely adopted controls baseline specifically applicable to OT environments and is an appropriate organizing framework for any security program maturity assessment in this sector.

CSA Resource Alignment

The ZionSiphon disclosure illustrates threat dynamics that map directly to several active CSA research areas.

MAESTRO (Agentic AI Threat Modeling Framework) provides a layered model for analyzing threats involving AI agents and their interaction with broader infrastructure. ZionSiphon's MCP-based C2 server — in which an LLM agent receives reconnaissance data, performs analysis, and generates operator-directed intelligence — maps to MAESTRO's concerns about agent execution context (Layer 3), deployment and infrastructure security (Layer 4), and the integrity of agent-to-human communication channels (Layer 7). The MAESTRO framework's guidance on trust boundaries between AI agents and human operators is directly applicable to understanding how defenders can detect and disrupt AI-augmented malware C2 architectures. [10]

AI Controls Matrix (AICM) extends the Cloud Controls Matrix to cover AI systems specifically. Water sector operators deploying AI-assisted monitoring and anomaly detection tools — increasingly recommended as a countermeasure against the kind of adaptive attacks ZionSiphon represents — should apply AICM controls governing AI pipeline integrity, model input validation, and the security of AI inference infrastructure. A compromised or adversarially influenced AI monitoring system in a water treatment facility is itself a critical vulnerability. [11]

CSA AI Organizational Responsibilities publications address governance structures for AI deployment across operations. As water utilities begin adopting AI-assisted threat detection, the governance and accountability frameworks described in these publications provide the organizational scaffolding necessary to manage AI tools responsibly and ensure human oversight is maintained for decisions with physical-safety consequences. [12]

Zero Trust principles apply with particular force to the IT/OT boundary that ZionSiphon crosses. CSA's Zero Trust guidance emphasizes continuous verification of device identity and network communications rather than perimeter-based trust assumptions. In OT environments where devices have long operational

lifetimes and change management is infrequent, applying micro-segmentation and protocol-specific allowlisting to ICS communications implements the "never trust, always verify" posture that would limit ZionSiphon's ability to reach Modbus, DNP3, or S7comm endpoints even after gaining a foothold on a corporate network host.

References

[1] Darktrace. "Inside ZionSiphon: Darktrace's Analysis of OT Malware Targeting Israeli Water Systems." Darktrace Blog, April 2026.

[2] Bleeping Computer. "ZionSiphon Malware Designed to Sabotage Water Treatment Systems." Bleeping Computer, April 2026.

[3] The Hacker News. "Researchers Detect ZionSiphon Malware Targeting Israeli Water, Desalination OT Systems." The Hacker News, April 2026.

[4] Infosecurity Magazine. "ZionSiphon Malware Targets Water Infrastructure Systems." Infosecurity Magazine, April 2026.

[5] SecurityWeek. "ZionSiphon Malware Targets ICS in Water Facilities." SecurityWeek, April 2026.

[6] Darktrace. "AI/LLM-Generated Malware Used to Exploit React2Shell." Darktrace Blog, 2026.

[7] CISA. "IRGC-Affiliated Cyber Actors Exploit PLCs in Multiple Sectors, Including U.S. Water and Wastewater Systems Facilities." CISA Advisory AA23-335A, December 2023.

[8] CISA, FBI, NSA, EPA, DOE, CNMF. "Iranian-Affiliated Cyber Actors Exploit Programmable Logic Controllers Across U.S. Critical Infrastructure." CISA Advisory AA26-097A, April 7, 2026.

[9] Industrial Cyber. "Waterfall Threat Report 2026 Finds Ransomware Slowdown Masks Deeper Shift Toward Nation-State Attacks on Critical Infrastructure." Industrial Cyber, 2026.

[10] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[11] Cloud Security Alliance. "AI Controls Matrix." CSA Research, July 2025.

[12] Cloud Security Alliance. "AI Organizational Responsibilities: Core Security Responsibilities." CSA AI Safety Initiative, 2024.

