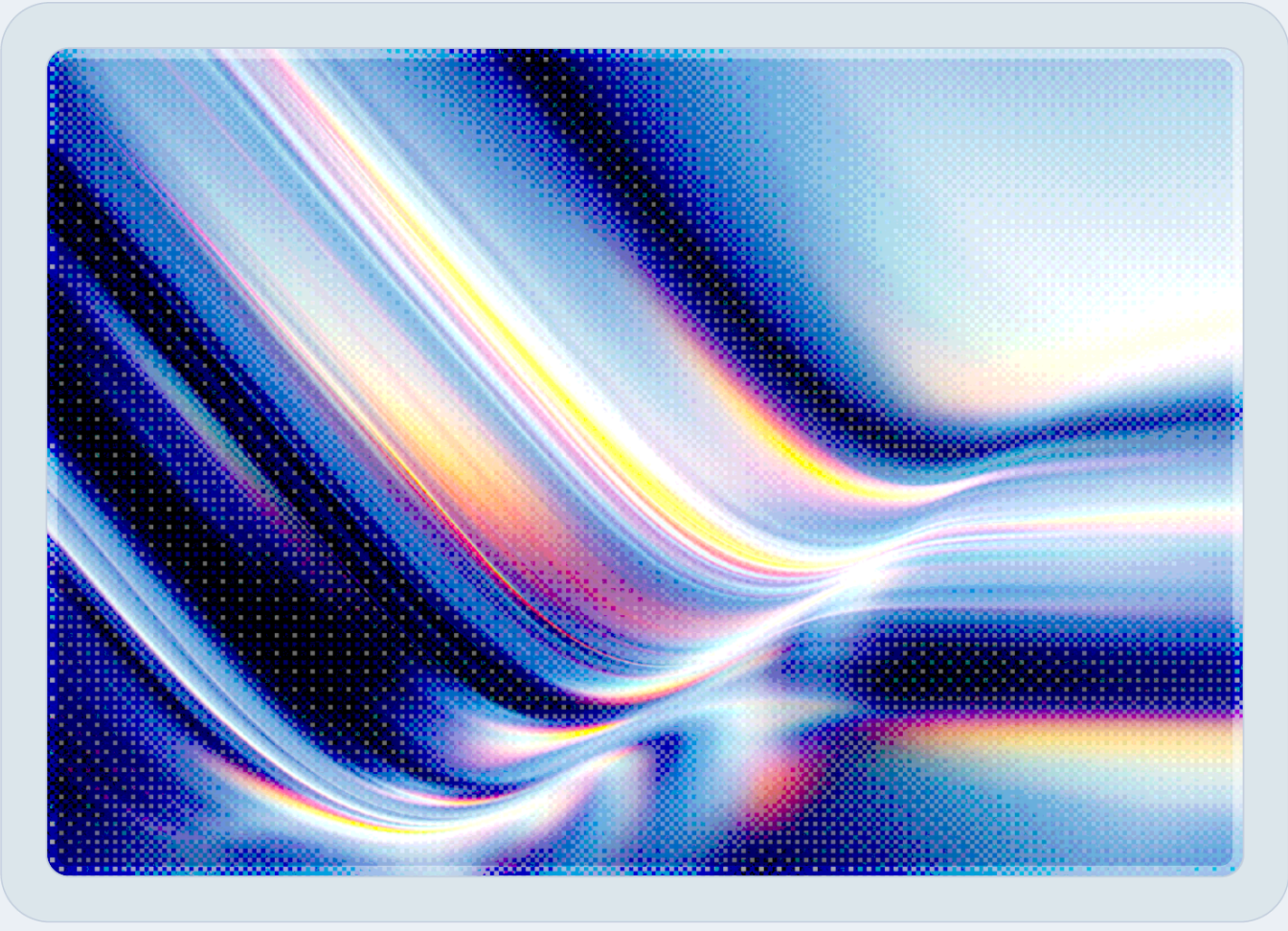
CSAI Foundation | Cloud Security Alliance

FortiBleed: Legacy Perimeter Devices as AI Agent Attack Vectors

How unpatched network infrastructure undermines enterprise AI security

2026-06-23

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- The FortiBleed campaign (June 2026) compromised credentials from tens of thousands of Fortinet FortiGate firewalls across 194 countries — credible estimates range from 74,000 to 86,644 depending on the enumeration window — demonstrating how deferred patching and outdated password-hashing practices convert perimeter devices into persistent, large-scale liabilities [1][2][5].
- A separately documented AI-augmented threat actor (January–February 2026) used a custom Model Context Protocol (MCP) server alongside commercial large language models to autonomously orchestrate FortiGate compromises across 55 countries in five weeks, illustrating that adversaries are adopting the same AI tooling enterprises are deploying [3].
- Compromised VPN and firewall appliances sit directly in the network path of AI agent infrastructure. Attackers with firewall-level access can intercept agent-to-tool communications, harvest machine credentials, and redirect traffic to attacker-controlled endpoints — transforming a perimeter breach into an agent orchestration compromise.
- CVE-2026-24858, a FortiCloud SSO authentication bypass rated CVSS 9.4, was exploited in the wild within days of its January 2026 disclosure and added to the CISA Known Exploited Vulnerabilities catalog with a three-day remediation deadline [4].
- Organizations deploying AI agents on enterprise networks must treat the underlying network infrastructure as a first-class AI security concern. A sophisticated AI agent running on a network bounded by compromised perimeter devices is not a secured system.

Background

Enterprise networks have long depended on VPN concentrators, next-generation firewalls, and SSL inspection appliances to enforce perimeter trust. Fortinet's FortiGate product line represents one of the most widely deployed examples: Shodan-enumerated estimates place tens of thousands of FortiGate management interfaces directly reachable from the public internet, making the platform a persistent target for credential-harvesting operations [2].

The unpatched-device problem is not new, but the proliferation of AI agent deployments on the same network segments raises the potential impact of each deferred patch. Fortinet's password-hashing practices changed materially with the release of FortiOS 7.2.11, 7.4.8, and 7.6.1, which introduced PBKDF2-based password storage to replace an older SHA-256 scheme that is computationally tractable on modern GPU hardware. However, devices that were upgraded from earlier firmware versions retain the weaker SHA-256 hashes for administrator accounts until each administrator next logs in and the system re-hashes their stored credential. An organization that applied firmware updates without subsequently cycling all administrator passwords is, in effect, still running SHA-256 credentials even on nominally current hardware [1][5].

This upgrade gap — the window between a software update and the complete realization of its security benefit — sits at the core of FortiBleed. Threat actors conducted systematic internet-wide scans of FortiGate management interfaces, exfiltrated device configuration files from reachable devices, and ran the extracted hashes against GPU-accelerated password-cracking infrastructure. The resulting credential database, organized by country, industry sector, and organization revenue, was shared or sold within threat-actor communities, yielding a verified credential set that researchers estimate covers approximately 50% of all internet-facing FortiGate appliances [2].

FortiBleed coincides with the period of most rapid AI agent deployment in enterprise environments. AI agents require network access — to APIs, tools, data stores, and external services — and they are typically deployed on the same enterprise segments that VPN and firewall appliances protect. This temporal convergence may not be coincidental: the same expansion of internet-connected enterprise infrastructure that created the FortiGate attack surface is now hosting the orchestration layers, tool servers, and MCP endpoints that AI agents traverse, compounding the potential blast radius of each perimeter compromise.

Security Analysis

The Credential Accumulation Problem

The FortiBleed dataset is distinguished not by a single novel vulnerability but by the accumulation of incremental operational gaps. Configuration files were reachable from the public internet because management interfaces had not been restricted to trusted administrative networks. Credentials were crackable because PBKDF2 migration had not been completed. The same credentials were valid across multiple devices because credential reuse practices had not been enforced. Each of these conditions is individually addressable through established operational security controls; their convergence at scale reflects the reality that deferred maintenance compounds across large device populations [1][5].

CVE-2026-24858 illustrates the parallel problem of unpatched software vulnerabilities. The flaw in Fortinet's FortiCloud SSO SAML implementation allows an authenticated FortiCloud account holder to bypass authentication and access devices associated with different accounts. CISA added the vulnerability to the Known Exploited Vulnerabilities catalog on January 27, 2026, one day after Fortinet's public disclosure, with an effective remediation deadline of January 30 — a three-day window reflecting active exploitation already underway [4]. Incident reports from multiple Fortinet customers described attackers creating rogue local administrator accounts on fully patched FortiOS devices, indicating that the SAML bypass was being chained with post-authentication configuration manipulation to establish persistence that survived firmware updates [4][6].

The CISA Known Exploited Vulnerabilities catalog has cataloged numerous Fortinet vulnerabilities exploited in active attacks, a significant share of which have been associated with ransomware campaigns [6]. This pattern indicates that Fortinet infrastructure is a recurring and reliably productive target rather than an occasional one. Organizations that treat each individual CVE as an isolated event rather than part of a sustained campaign against their device class are systematically underestimating their exposure.

AI-Augmented Attackers Target Network Perimeters

Between January 11 and February 18, 2026, a single Russian-speaking threat actor compromised more than 600 FortiGate devices across 55 countries in a campaign documented by Amazon Web Services [3]. What made the campaign notable was not the attack technique — the actor exploited exposed management ports and weak single-factor credentials, not a FortiGate vulnerability — but the AI-augmented operational infrastructure used to execute it.

The actor deployed a custom MCP server to ingest reconnaissance data and orchestrate attack workflows, and queried at least two commercial large language models (unnamed in public reporting) to generate step-by-step attack methodologies, expected success rates, and prioritized task trees for individual targets. A parallel port scanner handled internet-wide management port enumeration and credential testing across common FortiGate management ports. In some documented cases, an LLM client was configured to execute offensive tools autonomously — running Impacket scripts, Metasploit modules, and password-cracking utilities without operator approval for each individual action [3].

The operational implication is significant. The campaign's structure suggests a single operator delegated analytical and operational complexity to AI systems, enabling simultaneous intrusion operations across dozens of organizations in multiple countries without deep per-target technical expertise. Post-compromise activity included DCSync credential harvesting via Meterpreter and Mimikatz, pass-the-hash lateral movement, and targeting of Veeam backup infrastructure — a pattern associated with pre-

ransomware staging [3]. The campaign demonstrates that AI-powered attack tooling has crossed from theoretical capability to deployed operational infrastructure, and that automation substantially lowers the operational overhead previously required for sustained, multi-country intrusion campaigns.

The AI Agent Exposure Surface

While established frameworks such as MAESTRO and NIST Zero Trust principles identify the intersection of legacy perimeter compromise and enterprise AI agent deployment as a distinct threat surface, most enterprise AI deployments have not yet operationalized controls specifically addressing the perimeter-to-agent attack chain.

AI agents operate by traversing networks to reach tools, data sources, and external APIs. An agent orchestrated via an MCP server makes authenticated HTTP calls to tool endpoints, retrieves context from vector stores, and may interact with enterprise systems over protocols — LDAP, SMB, REST APIs — that have historically carried machine-to-machine trust. When the firewalls and VPN concentrators that bound this network are compromised, attackers gain several capabilities that are directly applicable to agent exploitation.

First, traffic interception. A threat actor with administrator access to a FortiGate appliance can configure SSL inspection, traffic mirroring, or policy-based redirection. AI agents communicating over HTTPS may be subject to man-in-the-middle interception if the firewall is trusted as a TLS termination point, allowing an attacker to observe API keys, session tokens, and the content of agent prompts and responses in plaintext [7].

Second, tool endpoint redirection. Agents that reach tool servers by hostname or internal DNS are vulnerable to DNS manipulation or BGP-level traffic steering initiated from a compromised perimeter device. Redirecting an agent's tool calls to an attacker-controlled endpoint enables prompt injection at the infrastructure layer, bypassing any agent-level input validation [8].

Third, non-human identity credential harvesting. AI agents authenticate to enterprise systems using API keys, OAuth tokens, and service account credentials. These credentials are transmitted over the same network paths that compromised firewalls can inspect. Once harvested, machine credentials are often longer-lived and broader in scope than human credentials; non-human identity lifecycle management has generally lagged user identity management across the industry.

The AI-augmented attacker campaign documented by AWS illustrates this bidirectionality: the same MCP-based infrastructure that defenders are adopting to give AI agents access to enterprise tools appeared in attacker tooling to orchestrate compromise operations against enterprise infrastructure. An

enterprise that has deployed MCP servers to manage AI agent tool access and has not secured the network perimeter around those servers has created a high-value target whose exploitation pattern is already documented in threat-actor campaigns from Q1 2026 [3].

Recommendations

Immediate Actions

Organizations running Fortinet FortiGate devices should take the following steps without delay. Expose no FortiGate management interface — web console, SSH, or FortiCloud access — to the public internet. Restrict management access to dedicated administrative networks, bastion hosts, or out-of-band management paths. Audit all devices against the FortiBleed credential database; several threat intelligence vendors have published lookups against device IP addresses or device identifiers from the leaked dataset [2][6].

For any device that has been upgraded from a FortiOS version predating 7.2.11/7.4.8/7.6.1 to a current version, all administrator accounts must log in after the upgrade to trigger PBKDF2 re-hashing. An upgrade without this step leaves SHA-256 hashes in place regardless of the installed firmware version [1]. Disable FortiCloud SSO on all devices unless the CVE-2026-24858 patch has been applied and a deliberate security review of SSO usage has been completed [4].

Short-Term Mitigations

Organizations that have deployed AI agents should audit the network topology those agents traverse. Identify which agents communicate over network segments protected by Fortinet or other edge appliances in the CISA Known Exploited Vulnerabilities catalog. Treat any network segment bounded by a device with an unpatched critical CVE as potentially compromised for purposes of AI agent trust decisions.

Implement mutual TLS authentication between AI agent runtimes and the tool servers or MCP endpoints they call. Certificate-pinned mTLS prevents traffic interception by an intermediate device, including a compromised firewall acting as a TLS termination proxy. Where mTLS is not feasible, scope API credentials to the minimum set of tool operations the agent requires, and rotate machine credentials on a schedule short enough to limit the usefulness of harvested credentials [3][8].

Adopt network segmentation that separates AI agent orchestration infrastructure — MCP servers, tool endpoints, vector stores — from general enterprise network segments. AI agent infrastructure should be treated as a high-value internal target, not as a generic internal server. Lateral movement from a compromised perimeter device to AI orchestration infrastructure should require crossing additional authentication and authorization boundaries [7].

Strategic Considerations

The FortiBleed campaign and the AI-augmented attacker campaign together illustrate a structural asymmetry in the current AI security posture of most enterprises. Defenders are deploying AI agents that require broad network access, while simultaneously maintaining perimeter infrastructure whose compromise rate is documented and whose patch cadence is lagging. Attackers are deploying AI-augmented tooling that exploits both simultaneously.

Addressing this asymmetry requires treating vulnerability management for network perimeter devices as a prerequisite for AI agent security, not a parallel track. An organization's assessment of its AI agent environment is incomplete without accounting for the security state of the infrastructure those agents traverse. Patch cadence for internet-facing devices should be governed by CISA KEV deadlines as a floor, not a ceiling, given that exploitation is frequently observed within days of public disclosure.

Longer term, organizations should evaluate Zero Trust architectures that eliminate implicit network-level trust from AI agent operations entirely. In a Zero Trust model, an AI agent's ability to reach a tool endpoint does not depend on the security state of the perimeter device that connects the agent's network segment to the tool server's segment — each call is individually authenticated and authorized regardless of network position. This architectural shift does not make legacy perimeter vulnerabilities irrelevant, but it significantly reduces the blast radius of a perimeter compromise for AI-specific workloads [8].

CSA Resource Alignment

The threat pattern described in this note maps directly to multiple layers of CSA's MAESTRO framework for agentic AI threat modeling. MAESTRO's infrastructure layer identifies the runtime environments, networking components, and orchestration platforms that support agent operation as distinct attack surfaces, and explicitly notes that infrastructure breaches can tamper with agent behavior by modifying tool endpoints or intercepting agent communications [9]. The FortiBleed and AI-augmented attacker campaigns provide concrete evidence that these threat vectors are active, not theoretical.

The AI Controls Matrix (AICM) addresses both the supply-side and demand-side of this risk. For AI customers and application providers, AICM controls governing network security, identity and access management for non-human identities, and incident response planning are directly applicable to the perimeter-to-agent attack chain described here. The AICM's coverage of the AI supply chain is also relevant, because the MCP-based tooling infrastructure that AI agents rely on is itself a supply chain component whose security depends on the integrity of the network through which it is reached [10].

CSA's Zero Trust guidance provides the architectural model for eliminating implicit network-layer trust from AI agent operations. The principle that no network position — including a position behind a nominally secure firewall — confers inherent trust applies with particular force when the securing device is itself in the CISA Known Exploited Vulnerabilities catalog [11].

The CSA STAR program's continuous monitoring tier offers a mechanism for organizations to communicate their posture on network perimeter patching and AI agent network security controls to customers and partners, supporting supply chain trust decisions in environments where AI agents are shared services consumed by multiple parties.

References

[1] Arctic Wolf. "Active FortiBleed Campaign Impacting Fortinet Devices Across 194 Countries." Arctic Wolf Blog, June 2026.

[2] SecurityWeek. "FortiBleed: 86,000 Fortinet Device Credentials Compromised." SecurityWeek, June 2026.

[3] Amazon Web Services. "AI-augmented threat actor accesses FortiGate devices at scale." AWS Security Blog, February 2026.

[4] CISA. "Fortinet Releases Guidance to Address Ongoing Exploitation of Authentication Bypass Vulnerability CVE-2026-24858." CISA, January 28, 2026.

[5] Help Net Security. "74,000 Fortinet firewall credentials exposed in FortiBleed data leak." Help Net Security, June 18, 2026.

[6] The Hacker News. "CISA Warns Fortinet Customers as FortiBleed Hits 86,644 FortiGate Devices." The Hacker News, June 2026.

[7] NIST. "Zero Trust Architecture." NIST Special Publication 800-207, August 2020.

[8] Kiteworks. "Agentic AI: Biggest Enterprise Security Threat for 2026." Kiteworks, 2026.

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

[10] Cloud Security Alliance. "Introductory Guidance to AICM." CSA Research, 2025.

[11] Cloud Security Alliance. "Zero Trust Advancement Center." CSA Research, 2025.

