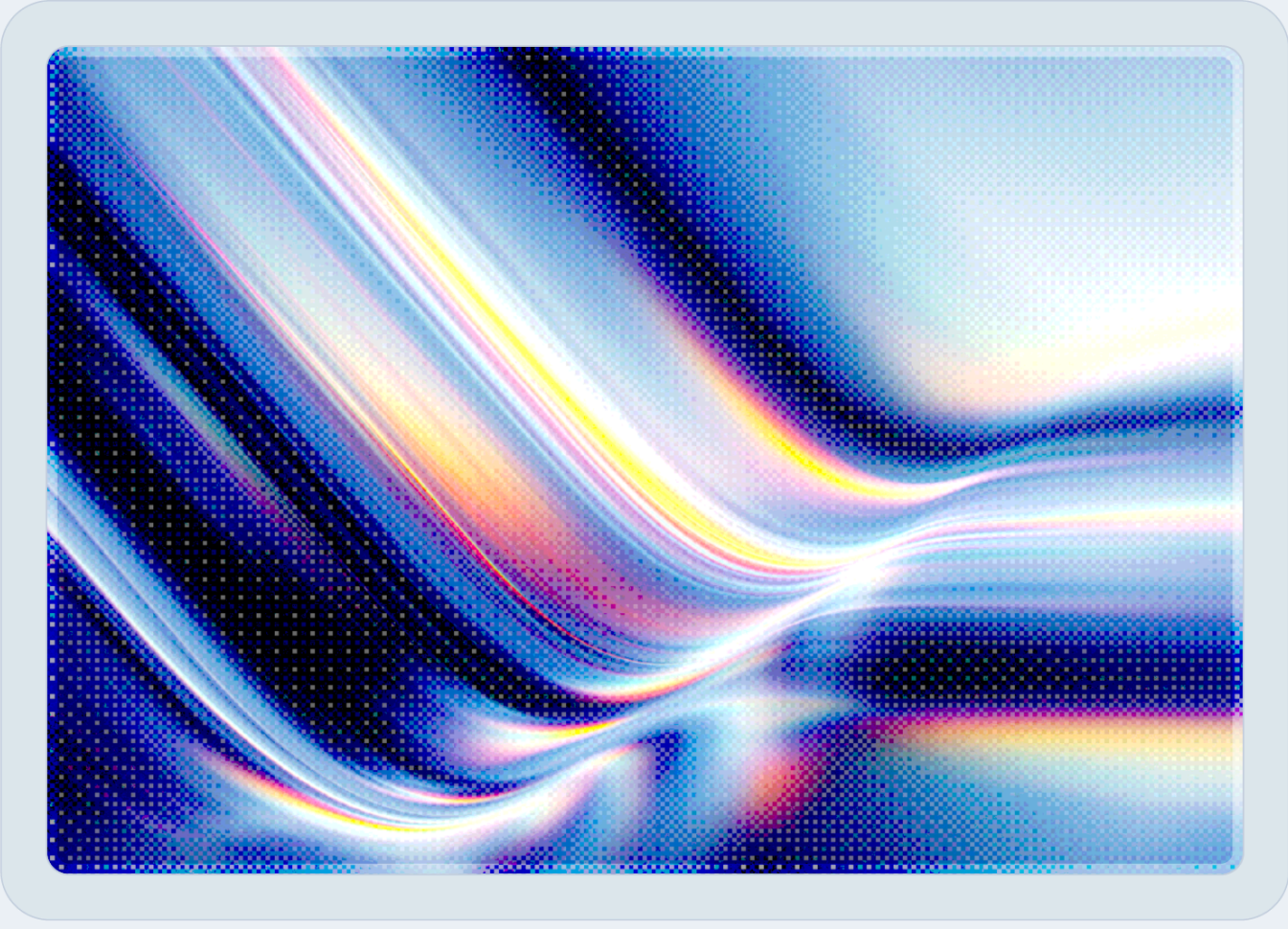
CSAI Foundation | Cloud Security Alliance

AI Finds 21 FFmpeg Zero-Days for $1,000

How Commodity AI Agents Are Collapsing the Economics of Vulnerability Discovery

2026-06-09

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Depthfirst, a security startup, used an autonomous AI agent built on commercially available Claude models to discover 21 previously unknown vulnerabilities in FFmpeg for approximately $1,000 in compute costs — roughly one-tenth the cost reported for comparable work by Anthropic's proprietary Mythos model [1][2].
- Several of the discovered bugs had been dormant for 15 to 23 years despite decades of continuous fuzzing and expert manual auditing by the security community [1].
- AI-driven vulnerability discovery has crossed from research demonstration to commodity capability: multiple independent systems — Google's Big Sleep, AISLE, TrendAI's ÆSIR, and now Depthfirst — are producing CVE-grade findings in production software at costs that were previously associated with a single day of consultant time [3][4][5][6].
- The economic bottleneck has inverted. Finding vulnerabilities is no longer the hard part. Triaging reports, shipping patches, and deploying fixes remains as slow as it has ever been, and AI-accelerated discovery is widening the window between identification and remediation [2].
- State-sponsored threat actors have developed AI-generated zero-day exploits for active operations, with Google's Threat Intelligence Group documenting the first confirmed AI-generated zero-day exploit staged for deployment in a planned intrusion campaign in May 2026 — disrupted by GTIG before mass exploitation could occur [9].
- Organizations relying on traditional patch cycle timelines — weekly or monthly — now face a structural mismatch with AI-accelerated discovery rates that has no near-term resolution.

Background

FFmpeg is among the most broadly deployed pieces of software that most users have never heard of. It is a free, open-source multimedia framework used to encode, decode, transcode, and stream audio and video, and it is embedded in every major web browser, in the infrastructure powering streaming platforms, in video conferencing clients, in content delivery networks, and in countless enterprise and consumer applications. Because it routinely parses complex, untrusted media inputs — streams from unknown sources, files of uncertain provenance — its parser and demuxer code represents a natural

attack surface for zero-click exploitation. A memory corruption vulnerability in FFmpeg's parsing logic, reachable through a maliciously crafted media file or RTSP stream, can enable remote code execution without any user interaction beyond opening the stream.

For most of the software's history, finding such vulnerabilities required substantial expert labor. Security researchers spent days or weeks tracing data flows through dense, pointer-heavy C code, developing hypotheses about boundary conditions, and constructing proof-of-concept inputs to confirm exploitation. Even with automated fuzzing tools like AFL and OSS-Fuzz, which have run continuously against FFmpeg for years, a meaningful proportion of memory-safety bugs have remained hidden in parsing paths that structured fuzz inputs rarely traverse. The economics of this process historically placed serious vulnerability research out of reach for most organizations. Industry estimates placed the cost of finding a single memory-safety vulnerability through expert manual review in the range of thousands of dollars, with some zero-days in hardened codebases commanding six-figure prices on exploit markets [9].

That economics model began shifting in 2025 as multiple independent AI research programs demonstrated capability at scale. Google's Big Sleep — a collaboration between Google DeepMind and Project Zero — announced in August 2025 that its AI agent had autonomously discovered 20 security vulnerabilities across open-source projects including FFmpeg and ImageMagick, with human experts reviewing findings before disclosure but no human involvement in the discovery itself [7]. The same year, AISLE's AI system began surfacing CVEs in OpenSSL at a rate that would account for 13 of the 14 OpenSSL vulnerabilities assigned across all of 2025 [10]. Anthropic's April 2026 announcement of Claude Mythos Preview — a model that found thousands of zero-days across every major operating system and browser, with individual Linux kernel exploit chains documented at under $1,000 each — marked an inflection point in public awareness of the trend [4]. TrendAI's ÆSIR platform disclosed 21 CVEs in AI infrastructure components including NVIDIA's stack, Tencent services, and MLflow, demonstrating that AI-assisted discovery was extending beyond general-purpose software into the specific systems underpinning enterprise AI deployments [6].

The Depthfirst disclosure of June 6, 2026 adds a new dimension to this picture: not the capability of a frontier model operated by a well-resourced AI lab, but the capability of a small startup operating commodity models on a startup budget.

Security Analysis

The Depthfirst Finding

Depthfirst's security research system analyzed approximately 1.5 million lines of FFmpeg's C code and produced 21 novel vulnerability findings at an aggregate compute cost of roughly $1,000 [1]. Critically, the system is built on publicly available Claude models — the research team explicitly noted they do not have access to Anthropic's proprietary Mythos model — placing this class of capability within reach of any organization or threat actor able to pay commercial API rates [1]. Anthropic had previously identified a 16-year-old H.264 flaw in FFmpeg using Mythos at a reported cost of approximately $10,000; Depthfirst achieved comparable scope at roughly one-tenth that cost [2].

The vulnerabilities are technically serious. Nine have been assigned CVE identifiers in the CVE-2026-39210 through CVE-2026-39218 range, all confirmed as heap or stack overflows in parsers and demuxers, and the remaining findings have been fixed upstream while pending CVE numbering [1]. The oldest confirmed finding, CVE-2026-39214, is a stack buffer overflow in FFmpeg's service-description-table handling code that has been present since 2003 — 23 years of continuous fuzzing, expert review, and security research produced no public disclosure [1]. A finding currently tracked only by Depthfirst's internal identifier DFVULN-127 is a heap buffer overflow in the AV1 RTP depacketizer reachable via a single 183-byte packet; the research report demonstrates instruction-pointer hijacking through heap corruption, reachable from an unauthenticated position simply by pointing FFmpeg at an attacker-controlled RTSP stream [1].

What distinguishes Depthfirst's approach from generalized AI coding tools is the architecture of its agentic security system. Rather than prompting a general model to find bugs, the system performs structured threat modeling to identify exposed attack surfaces, traces data flows through relevant code paths, generates parallel hypotheses about boundary conditions, and automatically produces reproducible proof-of-concept inputs before reporting any finding. The mandatory PoC validation step is significant: it eliminates the class of false-positive reports that caused FFmpeg maintainers to publicly characterize some AI-submitted bugs as "CVE slop" in the context of earlier disclosures [7]. Every Depthfirst finding is confirmed by execution before it reaches a disclosure workflow [1].

The Emerging Cost Curve

The Depthfirst result is part of a pattern visible across multiple independent research programs over the past twelve months. The following table summarizes confirmed cost data from major AI vulnerability discovery disclosures.

System	Codebase	Findings	Reported Cost	Source
Depthfirst	FFmpeg	21 zero-days	~$1,000	[1][2]
Anthropic Mythos	FFmpeg (H.264 flaw)	CVE-grade bugs	~$10,000	[4]
Anthropic Mythos	Linux kernel	RCE exploit chain	<$1,000 per chain	[4]
AISLE	OpenSSL 3.6.1	12/12 CVEs (Jan. 2026)	Not disclosed	[10]
WordPress AI pipeline	WordPress plugins	300+ bugs in 72 hrs	~$20 per finding	[8]
Google Big Sleep	FFmpeg, ImageMagick	20 findings	Not disclosed	[7]
TrendAI ÆSIR	NVIDIA/Tencent/MLflow	21 CVEs	Not disclosed	[6]

No two data points in this table come from the same system or methodology, and cost figures are not directly comparable — a $20 WordPress plugin bug and a $1,000 FFmpeg memory corruption finding represent different levels of difficulty and impact. What the table establishes is directional: across heterogeneous approaches, the emerging cost floor for AI-assisted vulnerability discovery is measured in hundreds to low thousands of dollars, not in the tens of thousands historically associated with expert manual research.

The WordPress finding illuminates how context shapes these numbers. The $20-per-vulnerability figure reflects specific characteristics of the WordPress plugin ecosystem — a vast number of plugins maintained by volunteers without dedicated security resources — and should not be generalized to hardened enterprise codebases [8]. FFmpeg, by contrast, is one of the most aggressively fuzzed open-source projects in existence. The fact that an AI agent surfaced 21 novel findings in it for $1,000 suggests the economic case for AI-assisted research extends well beyond low-hanging fruit.

The Inverted Bottleneck

The more consequential shift may be structural rather than purely technical. In the same week that Depthfirst disclosed its FFmpeg findings, Google shipped Chrome 149 with 429 security patches — a record single-release total, with over 100 classified as critical or high severity [3]. The two events in close timing illustrate the central tension organizations now face: AI systems are producing vulnerability findings faster than software maintainers, enterprise patch teams, and vulnerability management workflows can absorb them.

This tension was visible in the earlier Big Sleep disclosures. When Google's AI discovered a bug in FFmpeg's decoder for a 1995 video game codec, the FFmpeg project's response was frustration, not gratitude — the maintainers characterized the report as low-value AI-generated noise, noting the extraordinary overhead of triaging, reproducing, and patching a vulnerability in rarely exercised legacy code [7]. The curl bug bounty program, which ran from 2019 to 2026 and paid $90,000 across 81 genuine vulnerabilities over seven years, was effectively ended by the volume of low-quality AI-generated submissions overwhelming volunteer maintainers [10]. The disclosure process that took weeks per vulnerability was designed for the pre-AI discovery rate. It has not been redesigned for the post-AI rate.

This reframes which part of the vulnerability lifecycle is the limiting factor. Traditional vulnerability management concentrated effort heavily on the discovery phase — running scanners, subscribing to threat feeds, commissioning penetration tests. AI has made the discovery phase fast and cheap. The constraint is now the remediation phase: getting patches from upstream maintainers, testing them in staging, deploying them to production, and verifying coverage across affected systems. Organizations that have optimized for discovery but not for remediation throughput will find the gap between known vulnerabilities and deployed patches widening as AI-discovery velocity increases.

Threat Actor Implications

The same economics enabling defensive research also enable offensive campaigns. In May 2026, Google's Threat Intelligence Group documented the first confirmed instance of a threat actor developing an AI-generated zero-day exploit and staging it for deployment: a Python script targeting a web-based system administration tool, exploiting a two-factor authentication bypass — disrupted by GTIG before mass exploitation could occur [9]. Analysts identified the script as AI-generated based on stylistic hallmarks including educational docstrings and a hallucinated CVSS score embedded in the code itself [9]. Separately, PRC-nexus actors such as UNC2814 have been documented using AI for

vulnerability research on embedded devices, while North Korea's APT45 has independently been observed submitting thousands of repetitive prompts to AI systems for CVE analysis and proof-of-concept validation, accelerating the weaponization of known vulnerabilities [9].

The Depthfirst disclosure demonstrates that the same technical approach used defensively — an agentic system built on commercially available models, operating at commodity cost — is structurally available to any actor willing to pay API rates. The defensive researcher and the offensive adversary face the same commercial model menus. What differs is intent, accountability, and the legal constraints governing use. Security teams should plan on the assumption that adversaries with moderate technical sophistication are pursuing similar capability development, and that the cost barrier separating nation-state offensive programs from well-resourced criminal actors is lower than it has historically been.

Recommendations

Immediate Actions

Organizations running services or products that depend on FFmpeg should prioritize assessment and patching of affected versions in response to the Depthfirst disclosures. CVE-2026-39210 through CVE-2026-39218 are confirmed memory-safety vulnerabilities in FFmpeg's parser components; additional findings are documented in the upstream project without CVE assignment. The AV1 RTP depacketizer finding (DFVULN-127) warrants particular attention for any deployment that processes media from untrusted RTSP sources, as it requires no authentication and no user interaction beyond stream connection. Any deployment exposing FFmpeg's network streaming capabilities to untrusted sources should be treated as presenting active risk until the upstream patches are verified as deployed.

Short-Term Mitigations

The velocity at which AI systems are now surfacing vulnerabilities in foundational open-source software — FFmpeg, OpenSSL, curl, system kernels — suggests that organizations should not wait for CVE publication before acting on patch signals. Monitoring upstream project commit histories for security-relevant changes and deploying those changes before formal CVE assignment has historically been considered unusually aggressive practice. Given the current rate of AI-driven discovery, it is becoming practical necessity for high-risk components in critical environments.

Concurrently, organizations with significant media processing, streaming infrastructure, or AI workload exposure should evaluate AI-assisted security scanning for their own codebases and critical dependencies. The same agentic techniques Depthfirst applied to FFmpeg can be applied to internal

software. Vendor offerings built on similar architectures are now commercially available. The question of whether to use these tools defensively before adversaries use them offensively has largely been settled by events — the more actionable question is which codebases to prioritize and how to staff the triage function that follows.

Strategic Considerations

At the program level, organizations should begin redesigning vulnerability management workflows around the assumption that discovery velocity will continue to increase. This means investing in accelerated patch testing pipelines, automated regression testing frameworks that can validate security patches quickly, and escalation processes that can compress the standard patch cycle for critical components under active AI-driven research pressure. The traditional SLA-based approach — acknowledging a critical CVE within 48 hours, patching within 30 days — was calibrated to a world where critical CVEs were scarce and discovery was slow. In the current environment, that model produces structural lag between exposure and remediation [11].

Security leaders should also engage with the organizational dimension of the open-source maintainer bottleneck. The curl program's collapse under AI-generated submission volume illustrates that voluntary open-source disclosure infrastructure was not designed for the throughput that AI discovery generates. Enterprise consumers of open-source software have a direct interest in funding maintainer capacity for triage and patching — not only as a matter of ecosystem stewardship, but as a concrete input to their own patch timelines and SLA attainability.

CSA Resource Alignment

The Depthfirst FFmpeg disclosures and the broader shift in vulnerability discovery economics connect directly to several Cloud Security Alliance frameworks and initiatives.

CSA's **MAESTRO** threat modeling framework is applicable to the agentic AI systems now driving vulnerability discovery in both directions — defensive and offensive. The multi-layer threat model MAESTRO describes, spanning model providers, orchestration layers, and external services, applies both to organizations deploying AI security agents defensively and to characterizing threat actor AI pipelines. Security teams evaluating AI-assisted vulnerability scanning tools should apply MAESTRO threat modeling to the scanning infrastructure itself, which has privileged code access and surfaces sensitive findings requiring controlled handling.

The **AI Controls Matrix (AICM)** addresses vulnerability management through its software supply chain and application security control domains. Given the acceleration in AI-driven discovery across open-source infrastructure, the AICM's guidance on patch management timelines and dependency scanning should be evaluated against AI-discovery velocity benchmarks when establishing organizational control targets. The control gap between traditional SLA performance and AI-speed disclosure timelines is a measurable risk that AICM-mapped programs should quantify.

CSA's **STAR** (Security Trust Assurance and Risk) program provides a public disclosure mechanism for security practices. As AI-assisted vulnerability discovery becomes standard practice, STAR assessments for cloud providers and AI service operators should increasingly reflect whether organizations are applying AI-assisted security scanning to their own codebases, and what patch deployment timelines look like for AI-infrastructure dependencies including FFmpeg, OpenSSL, and media processing libraries.

The CSA AI Safety Initiative has previously analyzed Anthropic's Mythos announcement and its implications for the autonomous offensive threshold [12]. The Depthfirst result strengthens the central finding of that prior analysis: autonomous offensive capability at the zero-day level is no longer the exclusive province of nation-state actors or frontier AI labs. It is available at commercial API rates to any actor with the technical sophistication to build a focused agentic scaffold.

References

[1] Depthfirst. "21 Zero-Days in FFmpeg." depthfirst.com, June 2026.

[2] The Next Web. "An AI agent found 21 zero-days in FFmpeg for $1,000. Chrome just patched a record 429 bugs." thenextweb.com, June 6, 2026.

[3] The Hacker News. "AI Agent Uncovers 21 Zero-Days in FFmpeg; Chrome Patches Record 429 Bugs." thehackernews.com, June 2026.

[4] Anthropic. "Claude Mythos Preview." red.anthropic.com, April 2026.

[5] Help Net Security. "Anthropic's new AI model finds and exploits zero-days across every major OS and browser." helpnetsecurity.com, April 8, 2026.

[6] Trend Micro Research. "Introducing ÆSIR: Finding Zero-Day Vulnerabilities at the Speed of AI." trendmicro.com, 2026.

[7] TechCrunch. "Google says its AI-based bug hunter found 20 security vulnerabilities." techcrunch.com, August 4, 2025.

[8] Help Net Security. "$20 per zero-day is already the WordPress plugin reality." helpnetsecurity.com, May 22, 2026.

[9] Google Cloud. "Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access." cloud.google.com, May 2026.

[10] LessWrong. "AI found 12 of 12 OpenSSL zero-days (while curl cancelled its bug bounty)." lesswrong.com, 2026.

[11] Help Net Security. "The exploit gap is closing, and your patch cycle wasn't built for this." helpnetsecurity.com, April 15, 2026.

[12] Cloud Security Alliance AI Safety Initiative. "Claude Mythos and the AI Autonomous Offensive Threshold." labs.cloudsecurityalliance.org, April 2026.