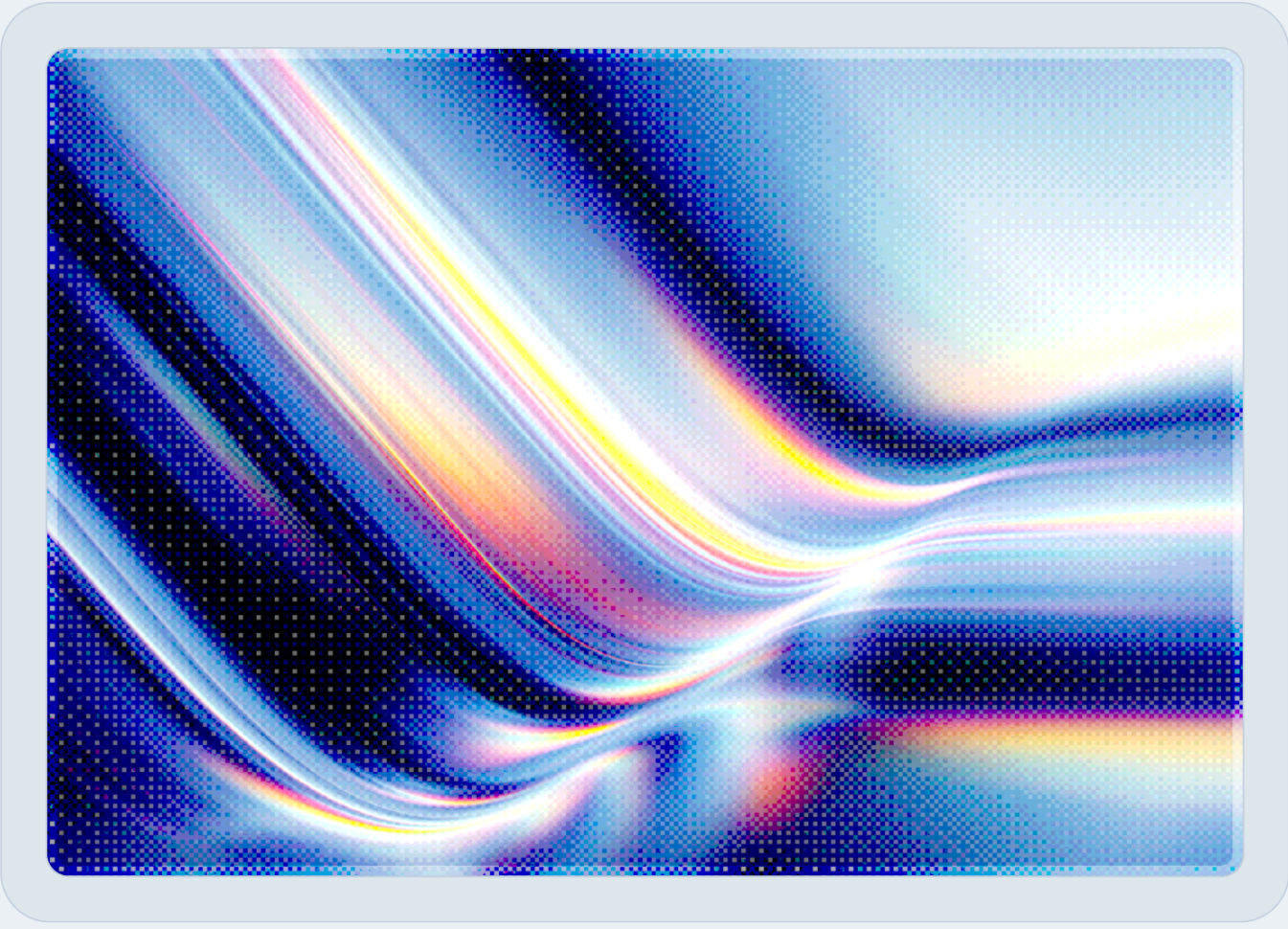
CSAI Foundation | Cloud Security Alliance

Spoofed and Disguised AI Agents: The Collapse of User-Agent Trust

Security Implications and Guidance for Web, API, and Commerce Operators

2026-11-13

AI-assisted Rapid Research



CSAI | cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

For three decades, web operators have treated the HTTP `User-Agent` header as a workable signal of who is calling, backed up when needed by published IP ranges and reverse DNS. The rapid growth of AI crawlers, AI assistants, and autonomous agents has broken that arrangement from two directions at once. Malicious automation now **spoofs** the declared identities of well-known AI agents to inherit the allowlists and rate-limit exemptions those agents enjoy. At the same time, some legitimate AI traffic is **indistinguishable from ordinary browsers**, whether through undeclared fetchers or because agentic browsers inherit a standard Chrome user agent, so it is neither blocked nor separately governed. The result is that a declared identity on its own now tells a defender very little about what is actually on the other end of the connection.

The first consequence is that any request claiming to be a named AI agent should be treated as unverified until its source network or a cryptographic signature confirms it. HUMAN's telemetry shows that a significant portion of requests claiming to be ChatGPT, Mistral, or Perplexity bots did not originate from those operators' infrastructure, although the report does not quantify the share [1]. Defenders should likewise assume that agentic browsers look like human Chrome users. ChatGPT Atlas and Perplexity Comet have been observed sending user-agent strings identical to desktop Chrome [2][3], and Cloudflare has reported that a declared crawler fell back to an undeclared fetcher with a generic Chrome user agent when blocked, a characterization the operator disputed [4][10].

Organizations should therefore begin piloting cryptographic agent authentication. HTTP Message Signatures (RFC 9421) and the IETF Web Bot Auth drafts [6] are currently the most widely referenced options: OpenAI signs its agent traffic with them [5], Akamai and Cloudflare offer verification [5][7], and Visa and Mastercard have built agent-commerce protocols on them [7]. Signature verification must nonetheless be kept separate from authorization, because a valid Web Bot Auth signature proves only that a request came from whoever controls a key directory URL; it does not prove that the operator is benign or that the request is authorized [6]. Finally, account, authentication, and checkout flows should be governed on the basis of behavior. HUMAN observed agentic traffic on account pages (8.8%), authentication flows (5%), and checkout pages (2.3%) in 2025 [1], and mistaken identity on these pages carries direct fraud and account-takeover consequences.

Background

The `User-Agent` request header has always been self-asserted. Any client can set it to any value, and browser vendors have long sent misleading strings for compatibility. The convention survived because the stakes were modest. Search engine crawlers published verification methods, site owners wrote `robots.txt` rules against declared bot names, and most bot management relied on network reputation and behavior rather than on the header itself. As long as the population of named automated agents was small and the agents gained little by misrepresenting themselves, self-declaration was good enough to steer most policy decisions.

Generative AI changed this in three ways. First, the volume of AI-driven traffic grew sharply. HUMAN reports 187% growth in monthly AI-driven traffic between January and December 2025, with three operators (OpenAI at 69%, Meta at 16%, and Anthropic at 11%) together accounting for roughly 96% of observed AI-driven traffic [1]. Fastly's analysis of January 2026 traffic found that bots made up 49% of all requests and that 99% of that bot traffic fell into Fastly's combined category of "unwanted or unverifiable," with 1% verified or wanted [8]. Second, AI agents now do more than read the web. They fill forms, manage accounts, and complete purchases on a user's behalf, so a misidentified agent can now cause more than a content-scraping problem. Third, many site owners began writing explicit policies that name AI agents, allowing user-initiated assistants such as `ChatGPT-User` while blocking training crawlers. These name-based allowlists create an incentive to impersonate the agents they favor, and the vendor telemetry discussed below is consistent with that incentive being exploited.

Operators have also blurred the line between "crawler" and "user" for their own products. Google states that its user-triggered fetchers "generally ignore robots.txt rules" because a user requested the fetch, and it advises site owners to verify such requests through reverse DNS and published IP-range files rather than through the user-agent string, which it notes can be spoofed [9]. Perplexity has made a similar argument that user-initiated assistant requests are not crawling [20]. Whatever the merits of that position, it means some AI traffic intentionally sits outside the robots.txt framework. Site owners then fall back on identity-based controls, and those controls are exactly what spoofing and disguise undermine.

Security Analysis

Two Failure Modes: Spoofing and Disguise

CSA's Zero Trust guidance distinguishes identity *spoofing*, which creates a false identity, from identity *abuse*, which misuses a legitimate one, because the two call for different detection and prevention controls [11]. AI agent traffic presents an analogous split, with a twist: the misleading identity can originate with an attacker or with a legitimate operator. Within the legitimate case, two situations should be kept apart. Undeclared fetchers diverge from an operator's own published identity, while agentic browsers may be indistinguishable from Chrome by design or simply by inheritance from Chromium, in the same way that many Chromium-based browsers send Chrome-matching strings for compatibility. The table below summarizes the three patterns.

Dimension	Spoofed agents	Undeclared fetchers	Agentic browsers
What is misrepresented	A malicious client claims to be a known AI agent (e.g., `ChatGPT-User`, `PerplexityBot`)	A real AI service fetches content under a generic browser identity and from addresses outside its published ranges	An AI agent driving a Chromium-based browser presents a user agent identical to desktop Chrome
Effect	Scrapers, fraud operators, and credential-stuffing or reconnaissance tooling inherit trust granted to named AI agents	Agent activity bypasses blocks applied to the declared crawler	Agent activity receives the default treatment given to human browser traffic
Trust exploited	Allowlists, robots.txt exceptions, and rate-limit exemptions granted to named AI agents	Default trust granted to unlabeled browser traffic	Default trust and analytics attribution granted to human browser traffic

Dimension	Spoofed agents	Undeclared fetchers	Agentic browsers
Documented examples	Requests claiming to be ChatGPT, Mistral, and Perplexity bots from outside operator infrastructure [1]	Reported Chrome-user-agent fallback when Perplexity's declared crawler was blocked (disputed) [4][10]	Atlas and Comet Chrome-identical user agents [2][3]
Primary detection approach	Verify source IP/reverse DNS or cryptographic signature against the operator's published data	Network-range and behavioral analysis; voluntary signing by the vendor	Behavioral and out-of-band signals; voluntary signing by the agent vendor

The spoofing problem is now documented in vendor telemetry. HUMAN's Satori threat intelligence team compared the declared identities of AI training crawlers against behavioral and infrastructure signals. It found that a significant portion of requests claiming to be ChatGPT, Mistral, and Perplexity bots did not originate from those operators' infrastructure, and it concluded that "organizations that allowlist crawler traffic based solely on user-agent strings are granting access to an unknown number of unauthorized actors" [1]. DataDome separately reported that it observed 7.9 billion AI agent requests across its network in January and February 2026, and that Meta's `meta-externalagent` and OpenAI's `ChatGPT-User` were the most frequently impersonated agent identities by request volume [12]. Fastly likewise warns that many unverifiable bots "attempt to disguise themselves as verified bots," so a declared identity such as ChatGPT may not reflect the client's true origin [8]. Because these figures come from different vendor networks using different methods, they are not directly comparable. Their consistent direction is the important signal.

The disguise problem is harder to address because the parties involved are often mainstream products, and because indistinguishability may result from ordinary browser compatibility practice rather than from intent. In August 2025, Cloudflare reported that when Perplexity's declared crawler was blocked, an undeclared fetcher using a generic macOS Chrome user agent and IP addresses outside Perplexity's published ranges continued to access content. Cloudflare estimated that this stealth traffic ran at 3 to 6 million requests per day across tens of thousands of domains, alongside 20 to 25 million daily requests from the declared agent [4]. Perplexity disputed Cloudflare's characterization [10]. Agentic browsers raise a related issue at a structural level. Simon Willison observed at launch that ChatGPT Atlas sent a user-agent string "identical to the user-agent I get for the latest Google Chrome on macOS" [2]. HUMAN

reports that simple user-agent checks will not reliably separate Atlas or Comet from Chrome, so detection has to rely on behavioral patterns and incidental signals [3]. The security concern lies less in a human browsing with these products than in their autonomous agent modes. For a defender, an AI agent driving a real browser engine from a residential or corporate IP address is very hard to tell apart from the human who launched it.

Why the Stakes Have Risen

Misidentification mattered less when automated traffic mostly read public pages. HUMAN's 2025 telemetry shows that 77% of observed agentic activity was on product and search pages, but a meaningful share reached account pages (8.8%), authentication flows (5%), and checkout (2.3%) [1]. At those points, the security question is no longer "is this a crawler that should respect robots.txt?" but "should this session be allowed to log in, change account state, or spend money?" An adversary who can pass as a trusted shopping agent may evade the friction that bot defenses apply to account takeover, carding, and fake-account creation. A legitimate agent that is misclassified as a bot, on the other hand, gets blocked, which breaks a customer-facing workflow. Cloudflare summarizes the merchant's dilemma directly: agent traffic has traditionally been classified by user agent and IP address, and "these fields can be spoofed, leading to inaccurate classifications" [7].

The same agentic tooling also appears as attack infrastructure. HUMAN reports that publicly exposed OpenClaw gateways were observed generating synthetic referral traffic with fabricated UTM parameters and conducting high-velocity directory brute-forcing against web applications [1]. This shows that general-purpose agent frameworks are already being used as attack infrastructure. Because such frameworks can be configured to present any user agent, an unverified declared identity provides little assurance about intent.

Legal and Governance Pressure

The disguise question has reached the courts. In November 2025, Amazon sued Perplexity, alleging among other things that the Comet agent concealed its automated nature by presenting itself as a Chrome browser session [13]. In March 2026, the U.S. District Court for the Northern District of California granted Amazon a preliminary injunction. The court found that Amazon was likely to succeed on its Computer Fraud and Abuse Act claim that Comet accessed password-protected accounts with user permission but without Amazon's authorization [14]. The concealment itself was Amazon's allegation, not a court finding [14]. In August 2026, the Ninth Circuit vacated the preliminary injunction and remanded the case for further proceedings, reasoning that an agent acting on a user's instructions reflects access

by that user [15][21]. For security teams, the practical lesson is that the legal status of user-delegated agents remains unsettled. In CSA's assessment, technical controls that identify and govern agents explicitly are likely to prove more durable than reliance on terms of service or litigation outcomes.

The Emerging Fix and Its Limits

A prominent industry response, backed by OpenAI, Cloudflare, Akamai, Visa, and Mastercard, is cryptographic agent identity built on HTTP Message Signatures (RFC 9421) [5][7][16]. Other approaches to agent identity and delegation also remain under discussion, and the field has not settled on a single standard. The IETF Web Bot Auth working group's protocol draft, currently at revision -01 (October 2026), lets an automated client sign each request and send a `Signature-Agent` header containing an HTTPS URL where its public keys are published as a JWK Set. A verifier fetches that directory, checks the signature, and can then log, rate-limit, allowlist, or block that URL as a stable identifier [6]. The draft is explicit about why this is needed, noting that "User-Agent, when used alone, can be spoofed" and that "IP blocks alone can present a confusing story" [6]. Adoption is already underway. OpenAI documents that its agent signs requests with RFC 9421 using `Signature-Agent: "https://chatgpt.com"` and a published key directory, and it notes that Akamai and Cloudflare can verify these signatures on a site's behalf [5]. Visa's Trusted Agent Protocol and Mastercard's Agent Pay both build on Web Bot Auth, adding tags that distinguish browsing (`agent-browser-auth`) from purchasing (`agent-payer-auth`) and a nonce to resist replay [7].

Defenders should be clear about what these mechanisms do *not* provide. The Web Bot Auth draft states that it "says nothing about who operates the Agent, whether the Agent is benign, or whether the request is authorized." It also warns that a resolvable key URL "does not say that the operator of that URL is honest, or that it is the same party everyone knows about" [6]. The draft does not authenticate human users, does not define delegation or revocation, and, depending on which components are signed, may allow a signature to be replayed against the same authority until it expires [6]. Signing is also voluntary. It solves the spoofing problem only for agents that choose to sign, and it does nothing about agents that look like browsers. Cryptographic identity therefore narrows the trust gap without closing it. It must sit alongside behavioral analysis, explicit authorization policy, and governance of what verified agents are allowed to do.

Recommendations

Immediate Actions

Organizations should start by auditing every place where a declared AI agent identity confers privileges. Common examples include WAF and CDN allowlists, robots.txt-driven exceptions, rate-limit exemptions, CAPTCHA bypasses, and analytics filters keyed on user-agent strings. Each such rule should require a second, harder-to-forge signal. For agents that publish IP ranges or reverse-DNS patterns, as Google and OpenAI do [5][9], verify the source network before granting the exemption. Where a CDN or bot-management provider offers managed verification of signed agents, enable it and log the verification result alongside the declared user agent. Finally, review authentication, account-management, and checkout endpoints to confirm that a claimed agent identity does not by itself reduce friction there.

Short-Term Mitigations

Over the next one to two quarters, security and platform teams should build an AI traffic inventory that tracks three things separately: declared identity, verified identity, and observed behavior. Mismatches between them, such as a `ChatGPT-User` claim from a residential proxy network, are strong candidate detection signals, consistent with the impersonation patterns vendors have reported [1][12]. Teams should pilot verification of RFC 9421 / Web Bot Auth signatures for the agents that matter to their business, starting with commerce agents that use the Visa and Mastercard profiles where relevant [7]. Policy should then key on the verified `Signature-Agent` URL rather than on the user-agent string. Because agentic browsers can resemble human Chrome sessions [2][3], sensitive flows should also apply behavioral and step-up controls, such as re-authentication, transaction confirmation, and velocity limits, that do not depend on correctly classifying the client. Organizations that operate their own outbound agents should sign their traffic, publish a key directory, and document their user agents and IP ranges, consistent with the norms Cloudflare has proposed for well-behaved crawlers [4].

Strategic Considerations

In the longer term, organizations should treat external AI agents as a distinct class of non-human identity that requires explicit authentication, authorization, and accountability, rather than as a variant of either "bot" or "user." That means defining in policy which agent operators are permitted, which actions verified agents may take, and how a user's delegation to an agent is evidenced and bounded. Teams should also expect an agent that holds a valid signature to be compromised, misconfigured, or misused at some point. Because the Web Bot Auth draft currently lacks revocation and delegation semantics [6],

organizations should track the IETF work and the commerce-network profiles and provide feedback through industry channels. Legal and risk teams should monitor the evolving case law on user-delegated agent access [14][15], because it will shape how much leverage site operators have to require agent disclosure. Finally, procurement teams should ask AI vendors whether their agents and agentic browsers identify themselves cryptographically, and should treat deliberate disguise, such as operating undeclared fetchers or refusing to offer a verifiable identity mode when asked, as a supplier-risk factor.

CSA Resource Alignment

The most directly relevant CSA work is *Agentic AI Identity and Access Management: A New Approach* [17]. It argues that traditional IAM protocols designed for humans and static machine identities do not fit autonomous, multi-agent systems. It proposes an architecture built on decentralized identifiers, verifiable credentials, an Agent Naming Service for discovery, and continuous, risk-adaptive trust scoring. The spoofing and disguise patterns described in this note are a concrete, present-day example of the gap that paper anticipates. Web Bot Auth supplies a basic, verifiable agent identifier at the HTTP layer. The paper's emphasis on attested capabilities, just-in-time credentials, and behavioral monitoring of agents describes the authorization and accountability layers that the IETF draft explicitly leaves out of scope.

CSA's Zero Trust Working Group publication *Using Zero Trust to Counter Identity Spoofing & Abuse* [11] provides the conceptual frame for this note. Its distinction between spoofing a false identity and abusing a legitimate one maps directly onto the spoofed-agent and disguised-agent patterns. Its principle that access decisions must rest on continually verified identity attributes and signals, rather than on asserted ones, is the core argument against user-agent-based allowlisting. Defenders can apply its detective and preventive control guidance to AI agent traffic with little adaptation.

For control mapping, the *AI Controls Matrix (AICM) v1.1* [18] gives organizations a structure for implementing these recommendations. Its Identity and Access Management domain covers authentication and authorization of non-human identities, and its logging, monitoring, and threat and vulnerability management controls support the declared-versus-verified-versus-observed inventory recommended above. Organizations threat-modeling their own agent deployments, or the external agents that interact with them, can also use CSA's MAESTRO agentic threat modeling framework [19] to place identity spoofing within the broader agent ecosystem and deployment layers.

References

[1] HUMAN Security. "The CISO's Guide to AI and Agentic Traffic." HUMAN Security, 2026.

[2] Simon Willison. "Introducing ChatGPT Atlas." Simon Willison's Weblog, October 21, 2025.

[3] HUMAN Security. "ChatGPT Atlas vs. Perplexity Comet: Agentic Browsers." HUMAN Security Blog, 2025.

[4] Cloudflare. "Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives." Cloudflare Blog, August 4, 2025.

[5] OpenAI. "ChatGPT agent allowlisting." OpenAI Help Center, accessed November 2026.

[6] T. Meunier and S. Major. "HTTP Message Signatures for automated traffic (draft-ietf-webbotauth-httpsig-protocol-01)." IETF Internet-Draft (work in progress), October 29, 2026.

[7] Cloudflare. "Securing agentic commerce: helping AI Agents transact with Visa and Mastercard." Cloudflare Blog, October 24, 2025.

[8] Fastly. "Nearly Half the Web Isn't Human: Inside Fastly's Threat Insight Report." Fastly Blog, April 16, 2026.

[9] Google. "Google's user-triggered fetchers." Google Search Central, accessed November 2026.

[10] TechRadar. "Perplexity accused of breaking a major online AI scraping rule — but it says it has done nothing wrong." TechRadar, August 2025.

[11] Cloud Security Alliance Zero Trust Working Group. "Using Zero Trust to Counter Identity Spoofing & Abuse." Cloud Security Alliance.

[12] DataDome. "Businesses Struggle to Identify AI Agent Traffic." DataDome Press Release, March 2026.

[13] Retail Dive. "Amazon sues Perplexity over AI shopping agents." Retail Dive, November 6, 2025.

[14] The Fashion Law. "Court Blocks Perplexity from Accessing Amazon in Agentic AI Lawsuit." The Fashion Law, March 11, 2026.

[15] Kezia Jungco. "Perplexity Keeps Its AI Shopping Agent on Amazon: What Happens Next?." eWeek, August 5, 2026.

[16] A. Backman, J. Richer, and M. Sporny. "RFC 9421: HTTP Message Signatures." IETF, February 2024.

[17] Cloud Security Alliance. "Agentic AI Identity and Access Management: A New Approach." Cloud Security Alliance, 2025.

[18] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance.

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

[20] Perplexity. "Agents or Bots? Making Sense of AI on the Open Web." Perplexity Blog, August 2025.

[21] Eric Goldman. "Ninth Circuit Lifts Restrictions on Agentic AI Accessing Amazon." Technology & Marketing Law Blog, August 2026.