The AI Identity Dilemma: Malicious Bots in Disguise

November 6, 2025

Radware's Cyber Threat Intelligence (CTI) team assesses with high confidence that malicious actors are actively developing and deploying bots that impersonate legitimate AI agents from providers like Google, OpenAI, Grok and Anthropic.

Key Insights

- Major AI platforms deployed interactive agent modes in 2024-2025: OpenAI's ChatGPT Agent (virtual browser), Google Gemini (real-time web interaction), Grok (agent mode), and Anthropic Claude (virtual browser) all rely on "good bots" that require POST request permissions for transactional capabilities, breaking traditional bot security assumptions.
- Radware's CTI team has identified a critical security gap in bot mitigation systems related to the emergence of AI agent modes from OpenAI, Google and Anthropic that now require POST request permissions.
- Malicious actors can exploit updated bot policies by spoofing AI agent identities to bypass detection systems, potentially executing large-scale account takeover (ATO) and financial fraud attacks.
- The cybersecurity industry is responding with cryptographic authentication standards (IETF Web Bot Auth), but implementation gaps create immediate risk for organizations.

Background

Although search engines still act as the central nodes of the internet, AI agents—the future of the web—are coming to replace them. With more and more users searching via AI services, we expect to see a drop in typical search engine crawler traffic and a sharp rise AI bot traffic which, unlike crawlers, interacts with the business logic of websites and applications.

Bot mitigation solutions have traditionally classified bots as "good" or "bad" using three primary parameters: User Agent (UA) verification, IP address validation against published ranges and restricting those good bots to GET-only requests.[1,2] This approach has been practical against traditional web scrapers and malicious crawlers.

However, the landscape fundamentally changed in 2024-2025 with the introduction of interactive AI agents. OpenAI launched ChatGPT Agent Mode in January 2025, featuring virtual browser capabilities, business system connectors and multi-agent orchestration.[3] Google Gemini introduced real-time web interaction with URL context tools supporting up to 20 URLs per request.[4] Anthropic Claude deployed Computer Use capabilities, enabling desktop interaction via mouse and keyboard.[5]

These AI agents require POST request permissions to execute interactive functions including booking hotels, purchasing tickets and completing transactions. OpenAI explicitly recommends that ChatGPT agents should not be limited to GET requests only, as POST requests are essential for their Responses API and tool-calling functionality.[6]

The Risk Factors

Radware CTI team identifies six risk factors that incentivize AI bot impersonations:

1. **Economic Pressure to Comply:** Businesses face powerful incentives to grant AI agents broad access to their websites and applications to remain visible in the next major channel for e-commerce and customer service. This creates pressure to weaken security controls for a new class of "trusted" automated clients.

2. **Static Verification Methods**: Bot mitigation systems still rely primarily on UA strings and IP ranges for "good bot" classification, methods that were acknowledged as fundamentally inadequate for modern threats.[7]

3. **JavaScript Rendering:** Legit AI bits can fully render dynamic web applications, giving them access to interactive components like login portals, account dashboards and checkout processes that are invisible to simpler bots. This means the gap between the traffic patterns of a legit bot and a malicious one has been minimized.

4. **POST Request Allowlisting**: To accommodate legitimate AI agents, security policies must now permit POST requests from entities identified as AI bots**.** This breaks a fundamental security assumption that good bots only read, never write.

5. **Spoofing Simplicity**: Attackers need only spoof ChatGPT's user agent and use residential proxies or IP spoofing techniques to be classified as a "good AI bot" with POST permissions.

6. **Expected Traffic Surge**: Application owners anticipate significant increases in legitimate AI agent traffic, creating a detection blind spot where malicious bots masquerading as AI agents are more likely to pass unnoticed by security and marketing teams monitoring anomalies.

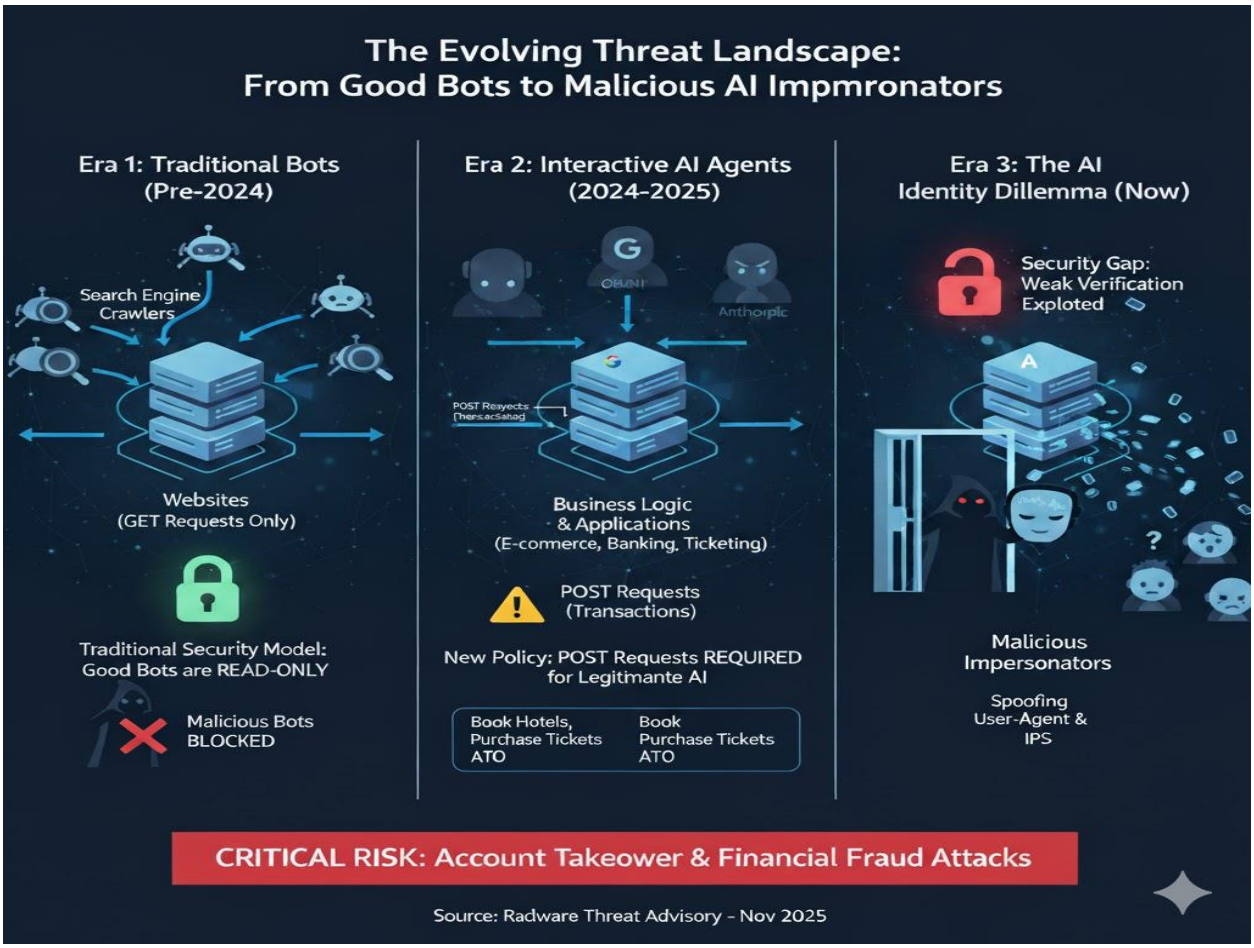

Figure 1: The evolution of threat actors from search engine crawler spoofing to AI bots spoofing (Source: Radware)

Target Risk Assessment

Organizations at highest risk include:

- **Financial Services:** Payment processors, banking platforms and cryptocurrency exchanges where ATO attacks have a direct monetary impact
- **E-commerce:** Online retail platforms where bot-driven purchases and inventory manipulation cause significant losses
- **Ticketing and Travel:** Event ticketing systems, airline bookings and hotel reservations are vulnerable to automated purchasing attacks
- **Healthcare:** Patient portals and telemedicine platforms where identity verification is critical

A Fractured Trust Model: AI Agent Identification

A primary factor enabling AI agent spoofing is the inconsistent landscape of identification and verification methods. An attacker can impersonate the agent with the weakest verification standard.

- **Google (Google-Extended)** leverages its existing infrastructure, recommending verification via a two-way DNS lookup or by checking against its published IP ranges—a strong method.
- **OpenAI (ChatGPT Agent)** has pioneered the gold standard: cryptographic HTTP Message Signatures (RFC 9421). Each request is signed, allowing for undeniable proof of origin that is resistant to all forms of spoofing**.** Other bots rely on published IP ranges.
- **Anthropic (ClaudeBot)** represents the weakest link. It relies solely on a User-Agent string for identification and does not publish official IP ranges, making it trivial to spoof.
- **Perplexity (PerplexityBot)** provides IP ranges for verification but controversially states its Perplexity-User agent intentionally ignores robots.txt rules, arguing it acts on a user's behalf.
- **Grok (xAI-Web-Crawler)** relies solely on a User-Agent string for identification and does not publish official IP ranges, making it trivial to spoof.

This fractured model creates a dilemma for security teams: implementing and maintaining distinct, provider-specific verification logic is a substantial effort, leading to exploitable security gaps.

AI Agent Identification & Verification Matrix

Provider	Agent Name	Verification Method(s)	Verification Strength	Notes
Google	`Google-Extended`	Reverse/Forward DNS Lookup, Published IPs	Strong	Used for Gemini model training.
OpenAI	`ChatGPT Agent`	Cryptographic HTTP Message Signatures	Very Strong	Gold standard; resistant to spoofing.
OpenAI	`ChatGPT-User`	Published IP Ranges	Moderate	Relies on IP allow-listing.
Anthropic	`ClaudeBot`	User-Agent String Only	Very Weak	No published IPs. Trivial to spoof.
Perplexity	`Perplexity-User`	Published IP Ranges	Moderate	Intentionally ignores robots.txt.
Grok	`xAI-Web-Crawler`	User-Agent String Only	Very Weak	No published IPs. Trivial to spoof.

Recommandations: A Security Team To-do List

1. **Adopt a Zero-trust Policy for State-changing Requests:** Any endpoint that accepts a POST request (e.g., login, registration, checkout) is a critical asset. Subject all automated clients attempting to access these endpoints to advanced, AI-resistant challenges like behavioral CAPTCHAs or proof-of-work checks.

2. **Treat User-Agent as Untrustworthy:** Any agent relying solely on a User-Agent string such as Anthropic's for identification must be treated as unverified by default

3. **Enforce Rigorous DNS and IP-Based Checks:** For all AIbots and especially Claude and Grok, security controls must perform two-way DNS lookups to verify the IP address matches the bot's claimed identity. If the legit bots have an IP range list, ensure your policy dynamically updates IP allow-lists from official sources.

4. **Prioritize Cryptographic Verification:** Implement and trust methods like OpenAI's HTTP Message Signatures as the highest-trust signal for agent identity.

5. **Prioritize BLA Defenses:** Legit AI bots and Malicious bots are both interacting with the target's business logic using headless browsers**.** It is crucial to move from static detection to dynamic detection that monitors not only how the bot looks, but also how it behaves and interacts with the application.

6. **Focus on Grok and Claude Spoofing:** These services don't share a specific IP range and therefore are easier to spoof.

Resources List

[1] NIST Technical Blog: Strengthening AI Agent Hijacking Evaluations (January 17, 2025)
[2] Cloudflare Blog: Forget IPs: using cryptography to verify bot and agent traffic
[3] OpenAI: ChatGPT agent - release notes (Note: The official announcement was in July 2025, not January)
[4] Google AI
    a. Gemini API Documentation - URL context
    b. Gemini API Documentation - Grounding with Google Search
[5] Anthropic Developer Documentation (Cookbooks & Guides)
[6] Anthropic Developer Documentation (Main)

EFFECTIVE WEB APPLICATION SECURITY ESSENTIALS

Full OWASP Top 10 coverage against defacements, injections, etc.

Low false-positive rate using negative and positive security models for maximum accuracy

Auto-policy generation capabilities for the broadest coverage with the lowest operational effort

Bot protection and device fingerprinting capabilities to overcome dynamic IP attacks and achieve improved bot detection and blocking

Securing APIs by filtering paths, understanding XML and JSON schemas for enforcement, and using activity tracking mechanisms to trace bots and guard internal resources

Flexible deployment options, including on-premises, out-of-path, virtual or cloud-based deployments

LEARN MORE AT RADWARE'S SECURITY RESEARCH CENTER.

To know more about today's attack vector landscape, understand the business impact of cyberattacks, or learn more about emerging attack types and tools, visit Radware's **Security Research Center**. Additionally, visit Radware's **Quarterly DDoS & Application Threat Analysis Center** for quarter-over-quarter analysis of DDoS and application attack activity based on data from Radware's cloud security services and threat intelligence.