POSITION PAPER

The New Fraud Frontier: How AI Agents Are Rewriting the Rules

AI agents are hitting your login and sign-up pages, and some of them are your partners - your payment processors, your customer service tools, your authorized integrations. But others are attackers—account takeover operations, sophisticated fraud rings. And here's the problem: they look exactly the same technically. The emergence of agentic AI has fundamentally disrupted traditional fraud and abuse detection, creating a landscape where beneficial and malicious automation are technologically indistinguishable. Tools like OpenAI Operator, Anthropic's Claude, Perplexity's Comet and browser-based agents like Opera Neon are rapidly evolving—creating both unprecedented opportunities and sophisticated new attack vectors that demand a complete rethinking of fraud prevention strategies.

When "Bot" Becomes a Spectrum

Not all AI agents are created equal, and understanding this nuance is critical. We at Arkose Labs analyze billions of sessions across our customers, and we are seeing three types of Agentic AI in the wild:

Good:

They self identify and perform actions that generate value for your business and the end user, like partner integration bots and agentic commerce bots.

Bad:

They are scaling fraud techniques and tactics like ATO, new account fraud (e.g. promotional credit fraud), payment fraud which previously required manual actions to evade bot detection. They do everything to masquerade as a legitimate user or bot and have a detrimental impact on your business.

Gray area:

They are helpful to the end users, but not so much to your platform or business like zero-click search bots. They operate without transparency but the level of harm is difficult to quantify.

How Agentic AI Disrupts the Status Quo

Legacy bot detection relied on assumptions that no longer hold. Behavioral patterns once distinguished bots from humans—but agentic AI now mimics human randomness with disturbing accuracy. Traditional bots followed rigid scripts; AI agents adapt strategies in real time based on the defenses they encounter. High-friction challenges were supposed to exhaust attackers, but AI agents have infinite patience. Traditional CAPTCHAs—the cornerstone of bot defense for decades—now achieve 80%+ solve rates by standard large language models.

Arkose's Dual-Track Defense Strategy

At Arkose Labs, we're approaching the agentic AI challenge through two parallel innovation tracks: AI Indicators of Fraud and AI-Resistant Mitigations.

AI Indicators of Fraud: Know Your Adversary

Rather than treating all traffic the same, our approach analyzes multiple dimensions:

API call patterns that reveal LLM agents making characteristic sequences (vision→reasoning→action)

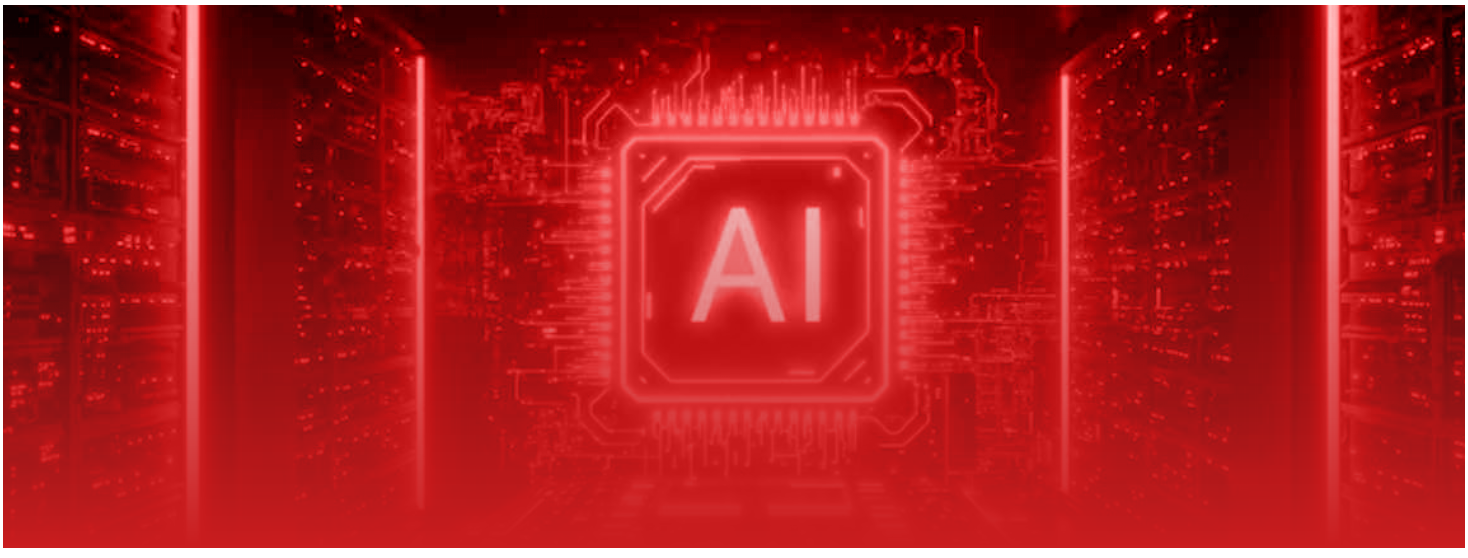
Framework detection through network timing analysis and WebDriver properties that expose AI wrappers

Proof of Work validation to assess if agents are masquerading and claiming to be a human user on a device. Agents run on cloud infrastructure, and will fail device checks.

Behavioral consistency analysis using statistical models to identify patterns that are "too perfect" or synthetically generated

Self-disclosure correlation examining user agent taxonomy, IP reputation, device intelligence and other signals to correlate self-disclosure with indicators of fraud.

The goal is to detect agentic AI even when they don't self-identify.



AI-Resistant Mitigations: Raising the Economic Bar

Our mitigation strategy focuses on making attacks economically unviable:

LLM-resistant challenges introduce new categories of visual puzzles specifically designed to be extremely difficult for current AI solvers.

Multimodal reasoning challenges require combining audio, visual and text elements—forcing attackers to chain multiple API calls across different foundational models, dramatically increasing their costs.

Next-generation proof-of-work scales difficulty dynamically based on the sophistication of detected solvers, with adaptive chaining that compounds computational requirements.

Behavioral biometrics perform cohort analysis of mouse movement, keyboard patterns and interaction timing to distinguish human variability from AI-generated behavior.

What We Are Seeing Today

We already have capabilities for detecting agentic AI sessions, and progressively applying appropriate AI resistant challenges to mitigate fraud and abuse. Using a combination of user-agent and signature-agent strings, and IP ranges/ASN's we can detect several self disclosing AI agents: OpenAI's GPT agent, Nova agent, Claude agent and Google Mariner agent. Based on co-occurring risk signals (infrastructure hiding, browser fingerprint manipulation etc.) our system can issue an appropriate LLM resistant challenge to collect additional signals and do progressive proofing of the type of agent.

Claude User Agent
Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; ClaudeBot/1.0; +claudebot@anthropic.com)

Over the last few months we have seen agentic AI traffic slowly increasing in the overall traffic we analyze, and some interesting trends are starting to stick out. The most common indicators of fraud and abuse we see in conjunction with agentic AI signals are:

Browser fingerprint manipulation: 74% spoofing browser values

OS impersonation: 69% faking macOS environments (likely to appear "legitimate")

Headless automation: 10% using Selenium/ headless browsers

Device fingerprint mismatches: Inconsistent WebGL, timezone and language signals

Behavioral biometrics (mouse movement) is another interesting attribute of agentic AI traffic. Most agents today have extremely short, precise and goal-oriented mouse movements. Arkose collects behavioral biometrics, and the following attributes can classify agentic AI traffic:

- **Movement Density.** The human session shows 15x more mouse events with continuous trajectories. AI Agents show sparse “Teleportation” behaviour.
- **Click Timing.** AI Agents are exhibiting superhuman precision with clicks faster than 10ms, impossible for human sessions due to hardware and human deficiencies.
- **Path Complexity.** Only 38% of the agentic events have unique positions vs 99% for the human sessions, indicating programmatic coordinate selection.

Beyond Detection: The Classification Imperative

The old question "is this a bot?" is dead. The new question is "is this agent authorized to do what it's trying to do?" This isn't about detecting automation anymore. It's about classification, management and governance at scale. Your good agents need to work. The bad ones need to be stopped. And the unknown ones need to be classified fast.

We combine economic disruption with intelligent classification, forcing attackers to burn through costly API calls while your legitimate automation flows seamlessly. As agentic AI continues evolving, the companies that will reap the business benefits while staying protected aren't those with the highest walls—they're the ones who know exactly who's at their door and can make that determination confidently and quickly.

BOOK A DEMO

Arkose Labs is the leading global provider offering a proactive fraud deterrence platform purpose-built to neutralize modern attacks, including those powered by Agentic AI and large language models (LLMs). Its comprehensive solution combines proprietary device identification (device ID), behavioral analysis, phishing protection, email intelligence, scraping prevention, API defense and bot management. Headquartered in San Mateo, California, the company maintains a global presence with offices throughout APAC, Central America, EMEA and South America.