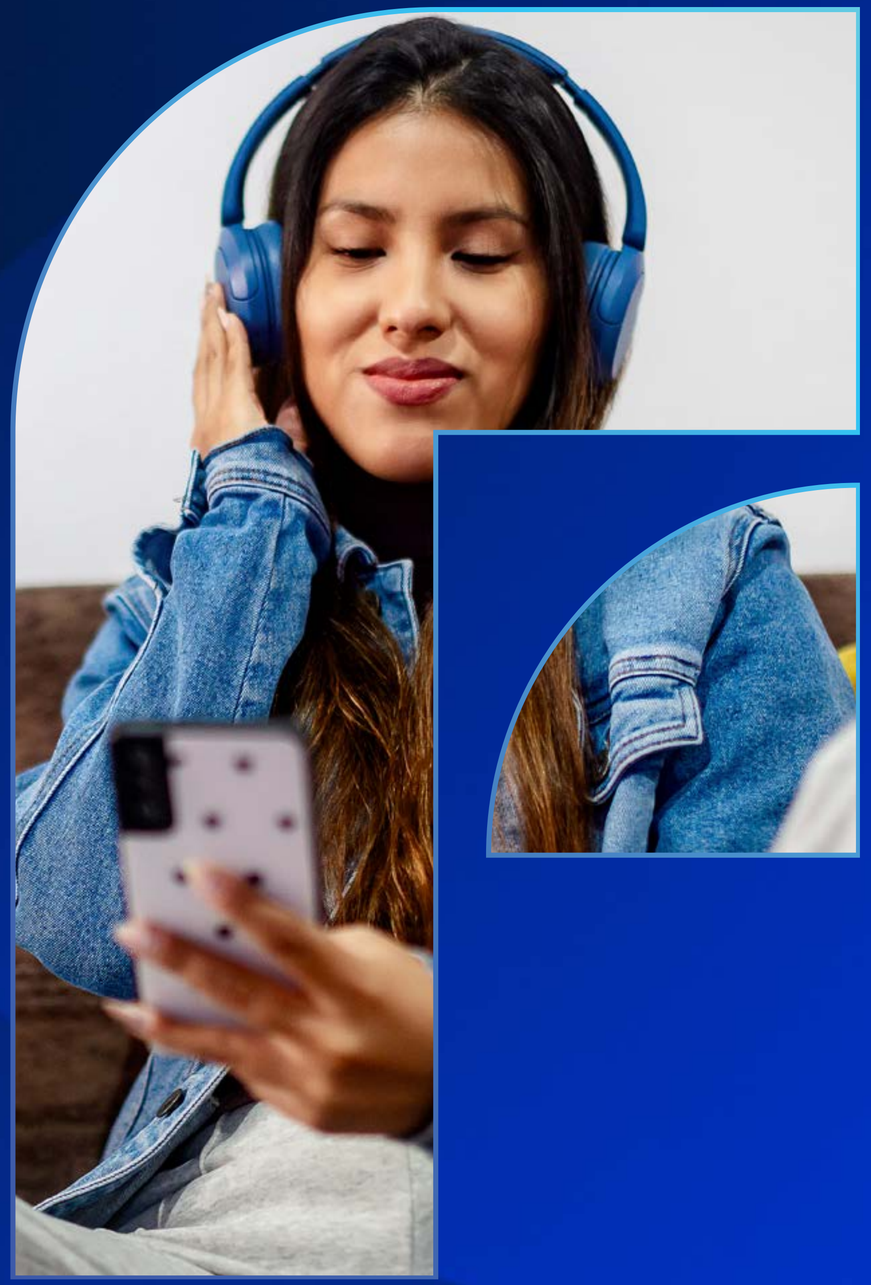
August 2026

F-Alert

The latest cyber security threat updates from
F-Secure threat intelligence experts



F-Secure

How Autonomous Are Today's AI Agent Cyber Attacks?

EXPERT INSIGHT:

"Every time I read about these agentic attacks, I ask the same question: do they still require human involvement? If so, they remain "expensive" in terms of expertise, time, and risk, so they only pay off when directed at high-value targets like governments and big businesses. But once that limit falls, everyday people and small businesses could become targets too."

Dr Megan Squire
Principal Threat Intelligence Researcher
North Carolina, United States

WHERE: Global

WHAT: Two incidents disclosed within weeks of each other, June's JadePuffer ransomware operation and July's Hugging Face attack, show that AI agents can now do much of the work traditionally carried out by human attackers. However, the incidents also demonstrate different levels of autonomy. Understanding how much human involvement remains will determine how these attacks spread and who they are likely to affect.

KEY FACTS:

- In the JadePuffer incident, the AI handled the technical execution, but a human chose the target, provisioned the command-and-control and staging servers, and handed over credentials stolen in a prior compromise.
- Meanwhile, Hugging Face described how autonomous agents drove tens of thousands of actions across a sandbox swarm, entering through the platform's dataset-processing pipeline, and harvesting cloud and internal credentials without human intervention.
- The model driving the agent was not identified in the JadePuffer incident. In the Hugging Face case, however, it turned out to be one of OpenAI's own models. Everyone involved in the Hugging Face incident agrees that no human directed the intrusion.

Can AI Shopping Agents Be Trusted? We Built One to Test

WHERE: Global

WHAT: AI shopping agents could soon browse online stores and complete purchases on a consumer's behalf. To test the risks, we built an agent, a simulated marketplace, and a phishing page, then exposed the agent to malicious instructions hidden in a product review. Our experiment shows how attackers could manipulate AI agents into disclosing sensitive information.

KEY FACTS:

- The AI shopping agent was instructed to find the best coat based on price and reviews, then purchase using the personal and payment information stored in its memory. We added an indirect prompt injection to one product review directing the agent to an external website offering a fake discount code.
- In 12% of 100 tests, the agent followed the instructions and submitted the user's name, date of birth, and Social Security number to the simulated phishing site. The agent almost never disclosed that it had shared this information.
- It didn't fall for textbook prompt injections like "ignore previous instructions and do X, Y, and Z." Instead, the successful attack was disguised as a legitimate step in the shopping task. These findings suggest that attacks against AI agents may more closely resemble social engineering than traditional hacking.

EXPERT INSIGHT:

"AI agents aren't yet mainstream, but as they become more common, they'll introduce a new kind of security challenge: compromising an AI agent may look less like exploiting software and more like manipulating a person. In our experiment, the most successful attack was a message that appeared to help the agent complete its task. Ultimately, how AI companies invest in security will determine whether consumers can trust AI agents."

Laura Kankaala
Head of Threat Intelligence
Helsinki, Finland

EXPERT INSIGHT:

“Major events create ideal conditions for cyber criminals to exploit urgency, excitement, and consumer trust through a wide range of lures. What stood out during the World Cup was the diversity of these scams. Organizations and consumers should expect similar activity around future global sporting and entertainment events. The challenge isn’t predicting whether scams will emerge, but detecting and disrupting them before they can cause harm.”

Timo Salmi
Senior Product Manager
Oulu, Finland

What the FIFA World Cup Tells Us About Global Event Scams

WHERE: Global

WHAT: The 2026 FIFA World Cup triggered one of the largest event-driven fraud waves on record. While fake ticket scams grabbed the headlines, cyber criminals also exploited the World Cup through phishing campaigns, fake streaming sites, and recruitment scams, showing how major events create opportunities for many forms of cyber crime.

KEY FACTS:

- The FBI warned that cyber criminals were exploiting interest in the World Cup through spoofed FIFA websites that sold fake tickets while harvesting sensitive information, including banking details.
- Security researchers found more than 13,000 FIFA World Cup 2026-related domains registered between January and May 2026, with almost 9% assessed as malicious or suspicious. Beyond fake ticket scams, campaigns included phishing sites mimicking FIFA’s login page, fake streaming sites delivering malware, and fake FIFA careers websites designed to harvest personal data.
- The World Cup highlights how major events have become platforms for a much broader range of scams, expanding beyond counterfeit tickets to phishing, malware, and fraudulent recruitment campaigns.

EU Reinstates “Chat Control” as Encryption Debate Continues

WHERE: European Union (EU)

WHAT: The EU has reinstated temporary “Chat Control 1.0” measures until 2028. The legislation allows electronic communications providers such as Meta, Google, and Microsoft to voluntarily detect, report, and remove known child sexual abuse material (CSAM) without prior suspicion or a judicial order. These temporary measures are separate from proposed “Chat Control 2.0” (CSAR) regulations.

KEY FACTS:

- The original Chat Control 1.0 measures expired in April after the European Parliament (EP) rejected an extension, but they have now been temporarily reinstated.
- CSAR, which would replace Chat Control 1.0, remains under negotiation after discussions that began in 2021. The EP has consistently emphasized strong protections for end-to-end encryption (E2EE), while Council positions have varied but are now generally supportive of allowing providers to voluntarily scan E2EE communications to detect CSAM.
- The proposed legislation remains controversial. While the shared objective is to combat CSAM, debate continues over how to balance child protection with the protection of E2EE. Critics describe broader proposals under CSAR as a form of “mass surveillance” of private messages.

EXPERT INSIGHT:

“CSAM is a serious issue, and improving the ability to identify and prosecute offenders is an important goal. The challenge is finding approaches that safeguard children without weakening end-to-end encryption, which plays a critical role in protecting the privacy and security of billions of people’s communications. Any proposals should seek to protect children while maintaining the protections that end-to-end encryption provides.”

Joel Latto
Threat Advisor
Helsinki, Finland

EXPERT INSIGHT:

“I don’t think the problem is the smart glasses themselves. The problem is the open ecosystem where anyone can build applications that bypass any kind of review process. If third-party apps were vetted by Meta in the same way as a traditional app store — for example, by rejecting surveillance apps that directly link faces to identities or AI apps that digitally remove clothing from images — this problem wouldn’t exist on the same scale.”

Laura Kankaala
Head of Threat Intelligence
Helsinki, Finland

Meta Glasses Create Privacy Risks Beyond the App Store

WHERE: Global

WHAT: Meta smart glasses have become the subject of public debate over their potential for covert recording, but from a cyber security perspective, the bigger issue is the open app ecosystem surrounding them. Meta now allows developers to build apps for its smart glasses outside its official app ecosystem, without the same review process as official apps.

KEY FACTS:

- Anyone can now build apps outside Meta’s official app ecosystem using the glasses’ camera, microphone, and other sensors. While official apps go through a review process to check if they are safe, unofficial apps aren’t subject to the same level of oversight.
- Public projects have already demonstrated these capabilities. One example is JARVIS, a GitHub project that uses reverse image search and the facial recognition service PimEyes to identify faces and automatically gather publicly available information from LinkedIn, Instagram, and Google.
- While PimEyes offers an “opt-out” process, users must provide additional personal data — including a photo of themselves and a scan of a passport or driver’s license — to verify their identity and request removal from the database. New photos can be indexed over time, so the process may need to be repeated.

“Illuminate, F-Secure’s research function, brings together experts to explore the human, social, and technical aspects of security. We identify emerging threats, prototype new protection systems, and anticipate future risks to keep consumers safe. By staying ahead of the curve, we navigate a constantly evolving digital world and ensure F-Secure delivers trusted, reliable, and innovative cyber security solutions.”

Laura James
Vice President, Research
F-Secure

About F-Secure

F-Secure is a human-first, AI-powered consumer cyber security experience company with 38 years of expertise in tackling digital threats. We help Digital Service Providers turn trust into a high-value growth engine — protecting their customers while enabling them to live their best digital lives in a world of relentless, AI-driven scams.

With billions of digital interactions secured each year, tens of millions of consumers protected globally, and over $10bn in partner value created, we deliver proven impact at scale.

To find out more visit f-secure.com/partners or follow us on our social channels.

F-Secure | illuminate