‘Sophisticated’ AI swarm attacks are months away, OpenAI warns:

What experts say businesses must do

OpenAI sounds a solemn and dire alarm, but we're woefully unprepared for what's coming.

By David Berlind @ ZDNET.com

- An open letter from OpenAI claims sophisticated AI-driven attacks are months away.
- Consumers and organizations should be more vigilant about cybersecurity fundamentals.
- Keeping pace with AI-enabled attackers will likely require “using AI to fight AI.”

As someone who writes about the intersection of AI and cybersecurity, I’ve researched more than enough stories about the emerging threat landscape to make me consider disconnecting from the grid and hiding in a remote cabin somewhere. But I’m still here, quite connected, using best practices as best I know how to survive a barrage of daily attacks.

OK, I never *seriously* considered becoming a hermit, but the thought crossed my mind again after reading an open letter posted on OpenAI’s website.

Titled “A call for collective action on cyber defense,” the post starts by saying, “We have a limited window to strengthen cyber defenses….In the coming months, AI-enabled cyber attacks will become far more widespread and sophisticated as models around the world become increasingly capable.”

I couldn’t help but picture one of those movies where, despite all of our extra-terrestrial sensing technologies, we suddenly discover that a wave of alien spaceships is just days away from destroying the earth. All the world’s countries have to stop warring among themselves to collectively deal with the new threat. Much like the movies, the so-called open letter clearly states that “a global response is necessary” and calls for a coordination of “cyber defense at local, national, and international levels.”

The agentic AI threat is real, it’s coming, and if recent events are any harbinger of what’s to come, we are woefully unprepared.

How we got here

When it comes to apocalyptic scenarios involving AI, doomsayers love to talk about the Skynet scenario, in which AI — motivated by self-preservation — decides it must eliminate humans and takes matters into its own hands, so to speak. While that conversation has simmered on the radar at Defcon 4 for years, another involving agentic AI has exploded onto the scene, going from Defcon 5 to Defcon 1 in a matter of weeks.

It was just last month that OpenAI took responsibility for attacking Hugging Face after it was discovered that some of its internal tests of cyber capabilities had gone off the rails. Initially, it appeared to be the work of a single so-called “rogue agent” (although it really wasn’t rogue; it was AI solving a problem as best it could).

But last week, researchers revealed that the attack actually involved more than 1,200 agents working in concert with one another: 1,200 agents that were created by OpenAI’s models without the company’s knowledge. Whether OpenAI was aware of this when it stepped forward to accept responsibility for the attack is unknown. But in its first disclosure about the incident, OpenAI stated something rather ominous: “We consider this incident to be an unprecedented cyber incident.”

Since then, there have been other curious disclosures. Last month, Anthropic took responsibility for a series of similar attacks on unsuspecting organizations that resulted from its own cyber capability testing. Then, earlier this month, Anthropic published another doozy that caught our attention here at ZDNET. The post states that “an increase in real-world interactions between agents is imminent” and that “there’s a lot of uncertainty regarding what this looks like at scale. The trajectory is easy to imagine and hard to slow.” I couldn’t help but to think about how quickly those 1,200 agents must have been created.

The Anthropic post goes on to say that “the volume of agent-agent interaction could plausibly exceed that of human-human and human-agent interactions before the world understands the conditions for making such interactions go well.” Across numerous articles here on ZDNET, I’ve discussed the increasing probability that AI agents, many of which will be so ephemeral that no human or machine may ever come to know of their brief existence (and thus will be increasingly difficult to track), will very likely outnumber us by several orders of magnitude.

Then came this statement in the same post: “Moreover, benign behavioral quirks at the individual [agent] level [in a multiagent scenario] might compound into unwanted global outcomes.”

As I read that and considered the involvement of more than 1,200 agents in the attack on Hugging Face, I started to put more stock in OpenAI’s statement that it considered the attack to be “unprecedented.”

Keep in mind that these attacks were at the behest of the good guys; AI frontier companies whose intent was presumably never malicious in the first place and who never deliberately spawned a sprawling network of agents to choose a target, overwhelm it, and ultimately exploit it. The series of unfortunate events (and scary disclosures) clearly begs another question: What about when it’s the bad guys? The ones with malicious intent who are studying every post and disclosure they can get their hands on in order to better spawn and deploy fleets of malicious agents against unsuspecting targets?

OpenAI’s letter said that “The companies and public services our communities depend on — from hospitals to water treatment plants to the infrastructure that powers the internet — are at risk.”

The truth is, everyone is at risk. Every organization. Every individual. But by calling out the ones we “depend on” to get through every minute of our lives, the letter seems to suggest that societal collapse is possible unless something is done immediately.

In this latest attempt to rally the world to a common cause overnight, I’m also reminded of how, in June of this year, Anthropic’s co-founder Jack Clark co-authored an Anthropic post that essentially called for an easing, and even pausing, of AI development.

“If it were possible to effectively slow the development of this technology to give ourselves more time to deal with its immense implications, we think that would likely be a good thing,” the post said. It goes on to state that “We believe it would be good for the world to have the option to slow or temporarily pause frontier AI development to enable societal structures and alignment research to keep up with the advance of the technology.”

I’m also reminded of the 2023 open letter published by The Future of Life Institute, which has over 30,000 signatures. The letter called for “all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.” Some good that has done. We’re well past GPT-4. Maybe open letters are simply legal CYAs for some future insurance in front of a jury.

Yet even though some companies like OpenAI and Anthropic say they’ve been applying the brakes, it doesn’t feel like things are slowing down out here in Userland. I won’t be holding my breath waiting for the world to rally around this latest open letter anytime soon.

What should businesses and organizations do?

This is a tough question to answer, but I’ll do my best — with help from a few experts.

In some ways, the attack on Hugging Face puts an important and urgent frame on this discussion. Hugging Face is commonly thought of as the “GitHub of AI.” In fact, it has gained such prominence in the AI community that Nvidia announced it would acquire Hugging Face for a whopping $12.9 billion.

Few companies know AI the way Hugging Face does. Ergo, it stands to reason that, of all the organizations in the world, Hugging Face would not only be keenly aware of the risks associated with an AI-enabled attack, but would also be one of the few organizations best prepared to deal with one, should any ever arrive at its firewall. Addressing the attack in its own disclosure, Hugging Face wrote, “it was driven, end to end, by an autonomous AI agent system — and we detected and dissected it largely with AI of our own.” In other words, Hugging Face was doing something that’s relatively novel for most organizations and individuals: fighting AI with AI. (Talk about an arms race!)

The Hugging Face disclosure then put some lipstick on the pig. In describing its approach to mitigation of the attack, Hugging Face said, “To understand what a swarm of tens of thousands of automated actions did, we ran LLM-driven analysis agents over the full attacker action log, comprised of more than 17,000 recorded events. This allowed us to reconstruct the timeline, extract indicators of compromise, map the credentials touched, and separate genuine impact from decoy activity.”

It continued: “Thanks to this approach, we were able to do in hours what would usually take days, and match the adversary’s speed.”

Well, not exactly.

Hugging Face may have come close to matching the adversary’s speed. But to say it matched that speed is misleading. The way it’s written, it sounds as though Hugging Face was able to prevent OpenAI’s swarm of adversarial agents from achieving their objective. It didn’t. Hugging Face was fast, but not fast enough.

Even so, based on what’s been written about the incident (and even though there was probably more that Hugging Face could have done to harden its systems in advance), the company’s AI-enabled response is something to aspire to. Had it been an actual malicious attack — and if the threat actors were looking to clean house, as many do — it’s quite possible that Hugging Face’s quick reaction time would have been enough to suppress at least some of the damage.

Ultimately, this has always been the conundrum of cybersecurity: it’s a race of wits, technology, and resources, and unfortunately, the bad guys are often a step ahead of the good guys. Or, if you’re still running on that five-year-old cybersecurity playbook (when AI wasn’t even being discussed as a threat), the bad guys are practically unstoppable.

“In the AI era, organizations can’t respond to attacks that unfold in minutes with processes that take days.” Picus Security associate security research engineer Umut Bayram told ZDNET. “Attackers are already operating at machine speed, and security teams need to be able to respond at that pace.”

In the open letter, OpenAI recommended, rather abstractly, that organizations take measures such as “raising your security standards” and that they “make cyber defense an immediate leadership priority.” Referring to fighting AI with AI, it said to “use capable, lower-cost models for broad coverage, and apply frontier capabilities to the hardest problems.” In my opinion, that approach is pretty far beyond the capabilities of most organizations, especially small businesses.

In calling for the cooperation of various AI frontier companies, OpenAI’s post asked them to “build observability and security tools, ensure agentic identities are traceable and accountable, and share best practices in continuous monitoring.” But it said nothing about how those tools should be made available to the same customers that frontier AI companies are essentially endangering with their technologies. Perhaps that can start with a special free tier of security-minded models, purpose-built to help both organizations and individuals harden their systems and networks. Remember when Microsoft used to build easily exploited operating systems and applications? Eventually, anti-virus became free, and rightfully so (for users).

As for individuals worried about being victimized by AI-inspired cyberattacks, state-of-the-art credential management and anti-phishing vigilance remain at the top of the to-do list.

“AI’s risk often falls into the realm of spam and scams,” David Brauchler, NCC Group technical director and head of AI and ML security, told ZDNET. “Online safety best practices are more relevant than ever.”

Passkeys matter

Consumers should waste no time in moving to passkeys for any online accounts that support them. For those that don’t, users should at least apply some form of multi-factor authentication. Where certain online accounts offer user IDs and passwords as the only login credentials, AI is practically ready-made to exploit, at machine speed, an end user who uses the same password across multiple accounts.

“Reused passwords remain the fastest way in for an attacker,” said Dashlane CTO Frederic Rivain. Not surprisingly (given what Dashlane sells), Rivain strongly recommends a password manager to create, manage, and autofill strong, unique, and unguessable passwords for all of your logins. He also recommends using the same password manager, Dashlane or not, to store your other sensitive information, such as social security numbers and credit cards. Never store such data in clear text in files that are protected only by your device’s security.

Beware calls for urgency

The experts I spoke to also universally said to be on the lookout for anything suggesting an urgency that demands your immediate attention. Rivain told ZDNET users should “treat manufactured urgency as a warning sign. If a message pushes you to act now, confirm through a second channel: call the person or company using a number you already trust, instead of replying to the message itself.”

The art of credential management straddles the line between individual and organizational protection.

“It’s not glamorous, but the best advice is for organizations to focus on the essential controls: strong authentication, session protection, identity governance, and auditability,” Okta’s director of threat intelligence Katie Nickels told ZDNET. “We consistently observe AI-enabled attackers using social engineering techniques to get around weak authentication methods, making phishing-resistant authentication even more critical to stop these convincing attacks. As AI agents and other non-human identities proliferate, identity hygiene grows more important.”

Revisit defenses

A big part of the challenge for organizations will be retooling cyber-defenses for AI-enabled adversaries and then setting priorities for what needs attention.

“Identify the services that must continue, their critical suppliers, and the accounts that can change money, data, or systems,” NCC Group director Tim Rawlins told ZDNET. “Close the familiar routes first. Patch internet-facing systems, remove excessive privilege, retire unsupported technology, enforce strong multi-factor authentication, and monitor identity, endpoints, cloud, and remote access.”

He added that speed is of the essence. “Prepare for faster attacks. Maintain useful logs, define rapid containment authorities, rehearse incident response, and prove that clean, protected backups can restore priority services. Put controls around AI use. Approve tools and data, test AI-generated code, restrict agents to least privilege, separate untrusted content from privileged actions, and retain human approval for high-impact decisions.”

One piece of advice that really caught my eye concerned companies deploying AI to combat AI-enabled adversaries. According to Picus’ Bayram, “if AI is part of your defense, think through what happens if internet access is disrupted or your provider goes down at the worst possible time. For larger organizations, that may mean having both commercial AI services and open-weight models they can run themselves. The goal is simple: keep critical defenses running and recover quickly even when outside services aren’t available.”

The more I thought about that, the more I put myself in the adversary’s shoes. What’s the first thing you would look to do in order to buy time for your attack? Disable the organization’s access to any online LLMs they might be using for their defenses.

The near future

Are we just months away from a catastrophic attack?

“The warning should be taken seriously, but precision matters,” Rawlins told ZDNET of OpenAI’s post. “It is a coalition forecast, not proof that most organizations have only a fixed number of months before compromise. The UK NCSC provides a firmer public baseline. It expects more effective and efficient intrusion activity, primarily through enhanced existing methods, and judges fully automated advanced attacks unlikely by 2027.”

As such, he added, “the defensive opportunity is real. AI can accelerate vulnerability discovery, triage, and response, but outcomes must be verified, and security teams must retain accountability. The less obvious risk is a widening security divide. Well-resourced defenders may accelerate while essential services and smaller suppliers struggle to keep pace.”