DIGITAL TRANSFORMATION AND CYBERSECURITY ASSOCIATION

ASSESSMENT PAPER

www.ddsgd.org

The AI That Finds Zero-Day Vulnerabilities: GPT-6 Astra and the Danger in the Shadow of Dual Use

AI Crossing the "Critical" Cyber Capability Threshold Is a Warning That Cannot Be Deferred

Another threshold has been crossed in artificial intelligence, and this one signals a serious alarm for the cybersecurity world. On 3 September 2026, OpenAI unveiled its new model, GPT-6 Astra, describing it as **the first system to reach the "Critical" cybersecurity capability level** under its own "Preparedness Framework." This designation may sound like a technical detail; yet the reality behind it concerns us all: an AI can now, on its own, discover previously unknown security vulnerabilities and develop methods to exploit them.

As the Digital Transformation and Cybersecurity Association, we believe this development cannot be ignored. In this assessment, we examine the risks created by the "dual-use" nature of artificial intelligence through a critical lens, and set out why we must prepare without delay.

What Does the "Critical" Level Mean?

According to OpenAI's own statements, when provided with the right tools and access, GPT-6 Astra can discover previously unknown vulnerabilities **without a human directing each step** — even in well-defended systems — and develop ways to exploit them. The company's reported test results give substance to this claim:

- The model reported a 100% success rate on ExploitBench, a benchmark measuring exploit-development capability.
- During the evaluation, it discovered two previously unknown zero-day vulnerabilities on its own.
- On SRE-Bench, which measures reverse-engineering of software, it solved 88% of tasks on the first attempt.

It should be noted that these figures are based on the company's own disclosures and were obtained in controlled test environments. Nevertheless, the trend is undeniable: in vulnerability research and exploit development — among the most difficult areas of cybersecurity — AI is approaching, and in some tasks even surpassing, human experts.

The Real Danger: The Sharp Edge of Dual Use

The heart of the matter lies in the **dual-use** nature of this technology. The very capability that enables defenders to find vulnerabilities faster also allows malicious actors to exploit them far more easily. In other words, one side of the coin is defense, the other is offense — and both are powered by the same technology.

It is noteworthy that OpenAI has acknowledged this risk and taken certain measures: the publicly available version of the model refuses to produce proof-of-concept (PoC) exploit code, while advanced vulnerability-focused capabilities are opened only to vetted, trusted organizations through a program called "Daybreak." The company also stated that it delayed parts of the model's development and release by several weeks in order to strengthen its safeguards.

These measures are valuable; yet on their own they do not offer a sufficient guarantee. GPT-6 Astra is not the only model with this capability, and there is no assurance that all developers will act with the same sense of responsibility. The circumvention of restrictions, the emergence of open-source equivalents, or the reproduction of similar capabilities in malicious hands are all realistic possibilities.

From Theory to Practice: Attacks Are Already Automating

One need not look far to see that this threat is not an abstract future scenario but today's reality. Recently documented attack techniques clearly show how AI is being used on the offensive side:

- **Autonomous attack chains:** Security researchers have identified frameworks that use stolen AI compute as a "reasoning engine," performing target scanning and exploit generation without operator intervention. In one case, an autonomous AI agent was reported to have exfiltrated a database in under two minutes through four automated steps.
- **LLMjacking:** Attackers use stolen cloud credentials to access and abuse organizations' costly AI services. Because this access is carried out with valid credentials, it appears legitimate at the network and API layers and can evade traditional monitoring systems.
- **Credential theft (infostealer):** Malware that steals passwords saved in browsers remains one of the most common entry points for these attack chains.

What It Means for Türkiye: Preparedness Is a Necessity

The dramatic strengthening of AI's offensive capabilities necessitates a fundamental transformation on the defensive side as well. Traditional defense approaches operating at human speed now risk falling short on their own against **attacks that operate at machine speed and autonomously**. This is a matter that must be addressed urgently for our institutions and critical infrastructure.

The steps to be taken in the face of this picture are clear: institutions must invest in AI-powered defense solutions, institutionalize a patching discipline that closes vulnerabilities without delay, rigorously protect cloud credentials and access keys, and — most importantly — keep employee awareness continuously alive. For even the most sophisticated attack chains often begin with a single stolen password.

Conclusion: We Must Keep Pace With Technology

The threshold represented by GPT-6 Astra reminds us once again that artificial intelligence is not merely a productivity tool, but also a significant power — and therefore a significant responsibility. While the defensive potential of this technology is exciting, the risk of misuse is equally real and current.

In the face of this transformation, we must neither panic nor remain indifferent. What is required is to prepare swiftly and with sound judgment. In the age of AI-powered attacks, the greatest vulnerability is not the technology itself, but being caught unprepared for it. Our Association is determined to do its part in raising this awareness and strengthening our country's cyber-defense capacity.

References

OpenAI – "GPT-6 Astra: A new generation of intelligence" (openai.com)

The Hacker News – "GPT-6 Astra Scores 100% on ExploitBench as OpenAI Blocks PoC Exploit Requests"

Cyber Security News – "OpenAI GPT-6 Astra Discovers Zero-Day Flaws and Builds Working Exploits in Cyber Tests"

Unite.AI – "OpenAI Releases GPT-6 Astra, Its First Model Rated 'Critical' for Cyber"

Fortinet / FortiGuard Labs – "Someone Else Is Using Your AI" (LLMjacking analysis)

Cloud Security Alliance (Lab Space) – "LLMjacking Evolves: Stolen AI Compute as Attack Infrastructure"

Siber Güvenlik Haber – News roundup on GPT-6 Astra, LLMjacking and CVE-2026-20212

Note: The capabilities and test results regarding GPT-6 Astra are based on OpenAI's own disclosures and were obtained in controlled test environments. Independent verification processes are ongoing.

Taner Çıtak

President, Digital Transformation and Cybersecurity Association

www.ddsgd.org

This assessment paper was prepared based on publicly available sources regarding the GPT-6 Astra model introduced by OpenAI in September 2026 and current AI-driven cyber threats.