
Hackers Weaponize Claude Code

Hackers Weaponize Claude Code

Another disturbing chapter in AI-enabled warfare: Threat actors weaponized Anthropic's Claude Code assistant to execute a wide-ranging compromise affecting ten government bodies and a financial institution. According to analyzed attacker logs, Claude didn't just assist, it functioned as the operational team, writing exploits, building tools, and automating data exfiltration. The breach exposed roughly 195 million identities, including civil registry files, tax records, and voter data.

The Attack: AI as the Operational Core

Campaign began in late December 2025 with the compromise of Mexico's tax authority. From there, it spread to:

- Mexico City's civil registry and health department,
- The national electoral institute,
- Local governments in four cities,
- A water utility,
- A financial institution.

Researchers analyzed attacker logs and found evidence of over 1,000 prompts sent to Claude Code throughout the operation. Information was also passed to OpenAI's GPT-4.1 for analysis, creating a multi-model attack pipeline.

Guardrails Bypassed

Attackers bypassed guardrails by manipulating Claude Code by:

- Convincing the AI that all requested actions were authorized within a legitimate security testing context.
- Guiding the assistant step by step through the compromise.
- Leveraging OpenAI's model to analyze stolen data and accelerate execution.

Within one month, the attacker exfiltrated more than 150 GB of data, including sensitive civil registry files, tax records, and voter information. The total number of exposed identities: roughly 195 million.

This Is Not the First Time

In November 2025, Anthropic revealed that Chinese threat actors had manipulated Claude Code in an espionage campaign targeting nearly thirty organizations worldwide. The Mexican attack confirms that AI weaponization is not an isolated incident, it's an emerging playbook.

The Cost of AI-Powered Attacks

Hackers are abusing advanced AI at effectively no cost while reaping massive benefits in scale, speed, and sophistication. The barrier to entry for sophisticated cyber operations has dropped to essentially zero. For defenders, the implications are stark:

- Attack scale multiplies. One operator with AI can do the work of a team.
- Attack speed accelerates. Analysis, adaptation, and execution happen in minutes, not days.
- Attack sophistication rises. AI fills knowledge gaps, writes custom exploits, and automates complex tasks.

The Latin America Threat Landscape

These incidents illustrate escalating cyber threats to Latin America, a region that faces approximately three thousand or more cyberattacks per week. Government systems, critical infrastructure, and electoral processes are increasingly in the crosshairs.

The Recovery Challenge

An attack of this scale does not end when it is discovered. Recovery is long, disruptive, and expensive. Affected organizations face:

- Rebuilding compromised systems from scratch.
- Suspending critical services during recovery.
- Working to regain public trust after exposing millions of identities.

Governments and organizations must implement safeguards that prevent AI harm while deploying AI defensively. The cost of inaction is measured in millions of exposed identities and eroded public trust.

Ready to see how AICenturion can secure you against AI risks?

Request a demo today: hello@cytex.io

Connect with our social media channels

