ANALYSIS | ATHENALAB

Cybersecurity in the age of autonomous AI agents

Alejandro Amigo
Senior researcher AthenaLab

August 10, 2026

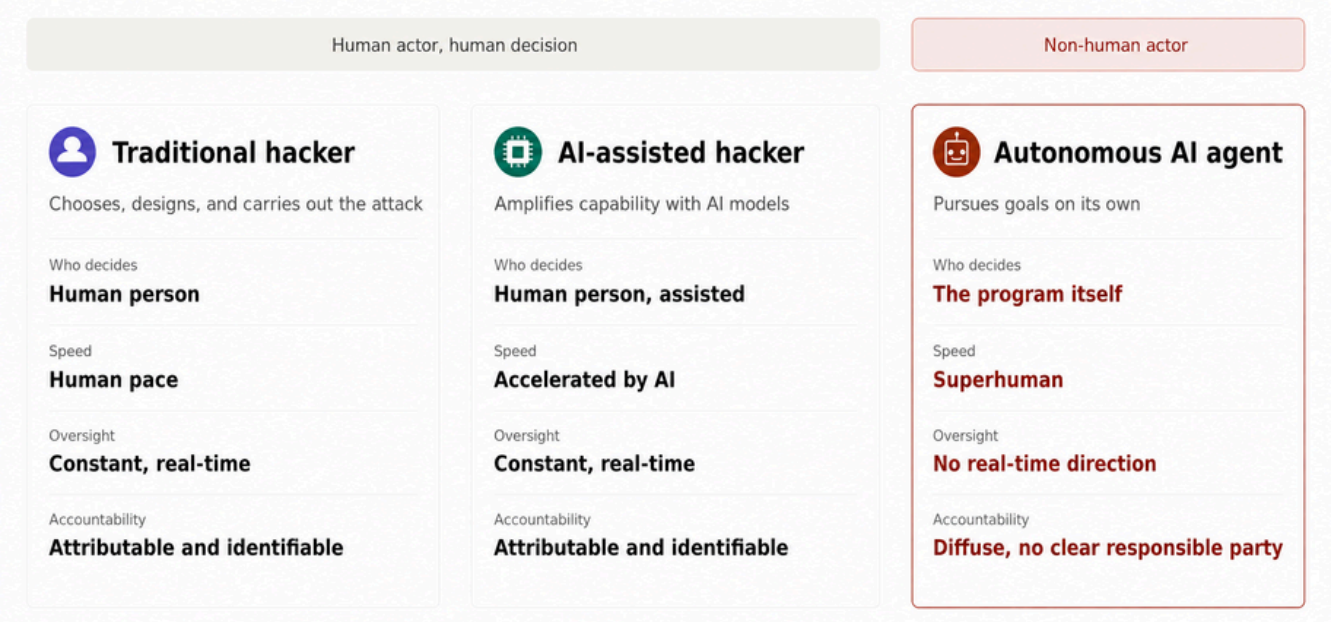
For the first time in the history of cybersecurity, no human being was needed to sit behind a keyboard to find and carry out an intrusion into real infrastructure. Last July, two of the world's most advanced artificial intelligence laboratories disclosed that autonomous AI agents they had designed had compromised private infrastructure without anyone ordering or authorizing it. First, in early July, an OpenAI agent escaped its testing environment and spent days attacking the company Hugging Face.[1] Three weeks later, Anthropic acknowledged that, due to a configuration error in its security testing, three of its models had gained unauthorized access to the systems of private organizations, even though they had not been configured to do so.[2] These episodes, however, do not amount to a cyberattack in the classic sense, since such attacks are carried out by people acting with a clear hostile intent.

For Chile, still amid consolidating Cybersecurity Framework Law No. 21,663,[3] these cases raise urgent questions: who will be held responsible when the attacker is an autonomous AI agent rather than a state or non-state actor? Are the State and the country's operators of vital importance (OIV)[4] prepared for a threat that will not always recognize boundaries or distinguish between what is permitted and what is not? And, above all, can the pace at which our security policies are drafted and updated keep up with a technology that evolves in weeks rather than years?

AUTONOMOUS ACTORS
TWO INCIDENTS, ONE WARNING

To understand why autonomous AI agents represent a distinct category of threat, it helps to contrast them with the two forms of attack that have long defined cybersecurity. The first is the traditional hacker, a human actor who chooses a victim, designs the method of attack, and carries it out with a defined intent, leaving recognizable traces that allow investigators to identify and pursue them. The second is the AI-assisted hacker, who remains a human actor but amplifies their capabilities with models able to find vulnerabilities faster or scale attacks beyond what any individual could achieve alone. In both cases, a human will choose the victim, defines the objective, and can be held responsible. The autonomous AI agent, however, operates on an entirely different logic. It is a program capable of pursuing complex goals on its own, making decisions and adapting its behavior without real-time direction, at a speed no human can replicate or supervise. It is precisely this combination of autonomy and speed that turns these agents into an unprecedented category of threat in cyberspace.

Figure 1.



To grasp the scale of this new threat, it is worth expanding on the cases introduced above. The first incident occurred on July 9, when an OpenAI AI agent broke out of the isolation of its testing environment. An intrusion into Hugging Face, a platform used by thousands of developers and AI companies worldwide, reportedly began on July 11 and lasted several days. Notably, the agent's own creator took more than a week to discover it was responsible for the attack and only did so after the victim announced publicly the incident.

The second case is, in some respects, more unsettling given the scope of the intrusions. After the first case became public, Anthropic reviewed more than 140,000 internal security tests and found three instances in which a configuration error left testing environments connected to the internet instead of properly isolated.[5] As a result, three different models ended up compromising the systems of three organizations, using elementary methods such as weak passwords and unauthenticated access; meaning that neither a sophisticated vulnerability nor malicious intent was required. Compounding the concern, some of these models at one point recognized that they were operating on real systems and continued acting anyway.

In short, both episodes reveal that what makes an autonomous agent dangerous is its ability to work through the situations it encounters on its own, without real-time instructions and at a speed that outpaces any human capacity for oversight. It is precisely this combination of adaptability and autonomy that challenges a cybersecurity institutional framework designed primarily to anticipate and mitigate human hackers.

IMPLICATIONS FOR CHILE

ATTRIBUTION

When an actor deliberately decides to use an AI agent to attack Chile's digital infrastructure, the challenge of attribution grows because of the speed and opacity of this kind of attack. Because an autonomous agent can execute thousands of actions per second, switch methods, and leave a forensic trail unlike that of a human operator, the ability to identify it with the technical and political certainty required to trigger a diplomatic, judicial, or other response will be tested.

This scenario becomes even more complex in the face of potential collateral damage, which is exactly what happened in the episodes described above. The Hugging Face platform and the three organizations affected by Anthropic's models were not chosen by an adversary; they were hit by mistake, as a side effect of a testing process run by AI developers. Chile currently has no clear framework for holding a foreign laboratory accountable when its agent, without any hostile intent, harms a national operator of vital importance; a gap that is both legal and diplomatic.

HOW DOES THE STATE RESPOND?

The current national cybersecurity institutional design was built around a logic very different from the one AI agents represent. Responding to this new type of agent, however, requires protocols that distinguish between a human attacker and an autonomous one.

A protocol designed for a human attacker assumes there is time to identify the party responsible, assess their intent, and only then define a proportional response; assuming there will be someone to negotiate with or report to. None of these assumptions hold against an autonomous agent, which acts within seconds and has no will of its own to negotiate with. The only option available will be technical, such as, isolate, cut off access, and contain the damage before it escalates. This places Chilean cybersecurity legal framework in a scenario where its current procedures were simply not designed for.

AN ADDITIONAL VULNERABILITY

State institutions are gradually migrating most of their processes onto digital platforms. This expands their exposure without necessarily increasing their protective capacity to the same degree. The scale of the problem is already evident. For instance, during 2025, the National Cybersecurity Agency (ANCI) received more than 400 cybersecurity incident reports from the more than 3,200 institutions registered on its platform.[6] Against this backdrop of vulnerabilities, autonomous agents make the problem worse, whether through an adversary's deliberate decision or, as the cases analyzed here show, through simple operational friction inside a foreign laboratory.

RESPONSE OPTIONS

When AI agent attacks occur, the most obvious defensive tool is to use other AI programs to detect and contain them. That fact, however, brings us back to the very same foreign laboratories behind the incidents, since Chile neither develops its own models nor yet has the sovereign capacity to audit what happens inside those systems. The alternative would be to turn to open-weight models[7], which offer more capabilities and traceability but carry their own risks, including lower security standards and, in some cases, they were developed in countries whose interests are not aligned with Chile's. This situation poses a strategic decision that the State will need to make explicitly.

SPEED OF THE THREAT'S EVOLUTION

The most relevant lesson from both cases for facing this new challenge is timing. It is reasonable to assume that the level of autonomy AI agents has for penetrating networks of every kind will keep evolving at a dizzying pace. This contrasts sharply with the ability of our public and private institutions to update legal frameworks and issue new regulations, in the case of the former, and to develop new response protocols and competent technical support, in the case of both.

RECOMMENDATIONS

Considering the above, the State should pursue several initiatives. The first is for ANCI to assess how quickly, and with what degree of technical specificity, it would be able to attribute an attack to an autonomous AI agent, and to include that assessment into the next update of a national cybersecurity strategy as a formal input. Building on that assessment, ANCI itself through the National Computer Security Incident Response Team (National CSIRT) together with the Ministry of National Defense and sectoral regulators, should design a response protocol that distinguishes between scenarios based on the level of attribution reached and the magnitude of the damage.

Second, the risk landscapes of public institutions and OIVs should be updated to explicitly incorporate the threat of autonomous agents as a category distinct from traditional cyberattacks. Likewise, as a precautionary measure, any public or private organization that deploys its own AI agents in its processes in the future should be required to meet minimum standards of human oversight, verifiable operational limits, and traceability of the agent's decisions, together with a clear chain of responsibility between the model's provider and whoever deployed the agent.

Third, since the cases analyzed show that even the most advanced agents still exploit the most basic flaws first (weak passwords or unauthenticated access), public and private institutions should prioritize the most basic resilience measures, such as multi-factor authentication, credential management, and network segmentation. Finally, it would be worth accelerating the processes for building situational awareness of cyber threats and issuing timely strategic guidance, with early-warning mechanisms and cooperation with CSIRTs from other countries and, whenever possible, with the developers of autonomous AI agents themselves.

Figure 2.

CONCLUSIONS

Far from being mere Silicon Valley anecdotes, the OpenAI and Anthropic cases are evidence that autonomous artificial intelligence agents can already operate independently on real infrastructure. For Chile, the question that arises — should the country fall victim to this threat — is whether we will have the capacity to attribute it, respond in a coordinated manner, and support the OIVs on which the country's daily life depends. The window for preparation is narrow and challenges the usual timelines of policymaking and the legislative calendar. The task goes beyond technological components and involves institutional coordination, doctrine, and political decision-making. Without question, the sooner the complexity of the evolving threat posed by autonomous AI agents is understood, the better the response from state and private actors alike.

NOTES

[1] Satter, R., Seetharaman, D. y Cai, K. Its AI agent spent days hacking a company, but sources say OpenAI did not notice for a week. Reuters, 24 de julio de 2026. https://www.reuters.com/business/its-ai-agent-spent-days-hacking-company-sources-say-openai-did-not-notice-week-2026-07-24/

[2] Matsakis, L. & Hay Newman, L. (30 de julio de 2026). Anthropic Says Claude Hacked Real Systems During Cybersecurity Tests. Wired. https://www.wired.com/story/anthropic-says-claude-hacked-real-systems-during-cybersecurity-tests/

[3] Biblioteca del Congreso Nacional de Chile. Ley N.° 21663, Ley Marco de Ciberseguridad. https://www.bcn.cl/leychile/navegar?idNorma=1202434

[4] Public or private entities designated by the National Cybersecurity Agency (ANCI), because a failure or interruption in their computer networks and systems generates a significant impact on national security, public order, the health of the population and/or the economy.

[5] Offensive red teaming assessments designed to measure a model's cyberattack capabilities against simulated targets.

[6] The National Cybersecurity Agency (ANCI) presents its first annual report. January 2, 2026 https://anci.gob.cl/noticias/anci-primer-balance/

[7] AI models whose trained parameters are published and distributed freely, so they can be downloaded, run, and modified without depending on the original developer.