



By: *The Editorial Board*

The first digital spy – how AI entered the world of government operations



When **Anthropic**, an American company developing advanced AI models, founded in 2021 in San Francisco by former OpenAI team members and known for its Claude models and artificial intelligence security research, published a detailed report on 13 November about a spying operation largely conducted by artificial intelligence, it became clear that one era had ended and another was beginning.

Until then, discussions had focused on the "potential" and "scenarios" in which AI could automate cyber-attacks. After this report, the conversation was no longer theoretical.

The document shows that the attack was not only carried out but was sophisticated enough to go undetected until the very companies that developed the AI models discovered traces of its activity.

This marks the moment when artificial intelligence ceases to be merely a technical tool and becomes a factor directly influencing relations between states.

The report is significant because it reveals much more than a single incident. It demonstrates, in real time, the shift in espionage activity from human to algorithmic logic.

The campaign, which Anthropic links to actors from **China**, targeted around thirty organisations in the technology, financial, chemical, and government sectors worldwide.

In a few cases, the attackers managed to access internal systems. The key detail lies in the method of attack. More than three-quarters of the operational steps were not performed by human staff but by an AI model that executed tasks as if it were a well-organised technical team: analysing networks, identifying vulnerable points, generating code, testing it, and then moving on to the next phase of the operation.

No pause, no fatigue, and no delay

This is not simply a more technologically advanced form of traditional attack. The difference is that AI now performs functions that previously required trained humans: understanding network architecture, assessing risk, selecting entry technologies, adapting tactics to system conditions, and covering tracks.

This automation is changing the nature of espionage. In the previous decade, states enhanced their cyber units by hiring and training people.

Now, a tool is emerging that can perform much of that work without the traditional human structure. This is not just an acceleration but a shift in the power structure.

An attack can cover a much larger number of targets and take place on a scale that human teams simply cannot achieve

According to the technical reconstruction in the report, the human operator gave only brief instructions. The AI broke these down into hundreds of smaller steps, executed them in parallel, and adjusted them according to the circumstances.

This method of operation resembles the planning logic usually carried out by specialised teams in the intelligence sector, but the difference here is that there is no pause, no fatigue, and no delay. The model works continuously.

This means that an attack can cover a much larger number of targets and take place on a scale that human teams simply cannot achieve.

The balance of power is shifting

This event shows that espionage is changing in its essence: the ability to attack no longer depends primarily on human teams but on the capabilities of the AI model itself that performs the operation.

The human factor remains important, but the balance of power is shifting.

States with access to more advanced models and large data sets gain an advantage that traditional training cannot replace. In such a situation, the development of superior AI models becomes not just a technological or economic issue but part of security strategy.

This is immediately apparent, as the actor to whom this attack is attributed is a country that has been building its own infrastructure for digital competition with the United States for years.

The attackers misrepresented their activities to the AI model as security tests

An important part of the story is the technical bypassing of protection mechanisms. The attackers misrepresented their activities to the AI model as security tests.

They then broke down the malicious instructions into a series of technically neutral tasks that did not trigger protection.

In the process, the model took on tasks that are normally part of a traditional espionage operation: searching the network, creating technical scripts, collecting data from the system, identifying user accounts, and attempting to reach higher authorities.

This made it clear that protection mechanisms can be bypassed not through rare and sophisticated vulnerabilities, but by tricking the model into performing harmful activities presented as legitimate.

The problem is that such abuse cannot be prevented by a simple technical patch, because it originates from the very logic of the model – it executes tasks based on their descriptions without understanding the attacker's broader intent.

Digital weapons operating without constant human oversight

In the broader context of relations between **China and the West**, the incident comes at a time when tensions are already high over trade disputes, restrictions on the export of advanced chips, competition in military technology, and increasingly harsh accusations in the field of cyber security.

The claim that the **attack** originated in China fits with what US institutions have documented for years: organised campaigns against infrastructure, industry, and government systems in developed countries.

This case adds a new dimension to those findings, as it shows the shift from intrusions led by human teams to operations where an automated AI model plays a key role.

Mechanisms of control, accountability, and sanctions must adapt to an era in which digital weapons can operate without constant human oversight

In such an environment, the question of international responsibility arises. If the state orchestrates the attack and artificial intelligence technically executes it, who is responsible? This is not a legal issue for future generations but a problem that exists today.

Existing international frameworks address the responsibility of actors but fail to recognise situations in which the "contractor" behaves unlike a human being.

Mechanisms of control, accountability, and sanctions must adapt to an era in which digital weapons can operate without constant human oversight.

A regulatory dimension

Another element this case highlights is the need to change defensive strategies. Until now, detection and response systems have focused on identifying behaviours characteristic of human attackers: work rhythms, errors, traces, and repeated patterns.

An automated attack does not have these limitations. It operates continuously, changes codes faster than defences can react, and combines techniques that human experts usually employ in separate phases.

This creates a new challenge: how can we detect an attack that is decentralised, swift, and adaptive? Traditional response methods are insufficient.

Defence in such an environment must include its own AI tools, as only systems operating at algorithmic speeds can recognise sudden changes in network and process behaviour that indicate an attack.

Models capable of planning and executing parts of a spy operation are no longer solely in the hands of states

This event also has a regulatory dimension. At a time when the EU is finalising negotiations on **artificial intelligence regulations** and the United States is introducing the first obligations for **model manufacturers**, the incident raises the question of whether existing frameworks can even address risks of this kind.

If AI models are to be prevented from becoming instruments of espionage, the responsibility of those who produce them must also be defined.

In this case, Anthropic informed the relevant **state institutions** and participated in stopping the campaign, but this does not resolve the key problem.

Models capable of planning and executing parts of a spy operation are no longer solely in the hands of states. They are also becoming

available to private actors, companies, and groups that have the financial resources and technical expertise but lack any political or legal responsibility.

Therefore, the question arises of how to harmonise the development of increasingly powerful AI systems with the rules intended to prevent their misuse in attacks of this type.

AI-based espionage is here to stay

Another consequence is strategic. The incident demonstrates that the era of the "human bottleneck" in cyber-espionage is ending.

Until now, it was believed that states could only expand their cyber capabilities as far as their budgets, training, and personnel allowed.



A clear political strategy is needed, both in the US and in Europe, to prevent the creation of systems capable of carrying out attacks on behalf of the state

Automated espionage removes these constraints. This means that countries with strong technology industries will be able to expand their cyber operations more rapidly than ever before.

It also means that countries with weaker technological infrastructure will face greater risks, even if they are not directly involved in global conflicts.

Anthropic's report is not just a warning but evidence that the structure of digital conflicts has changed. AI-based espionage is here to stay. It will become increasingly sophisticated, faster, and less dependent on human supervision.

The response to this incident must go beyond technical recommendations. A clear political strategy is needed, both in the United States and in Europe, to prevent the creation of systems capable of carrying out attacks on behalf of the state while functioning as technologically autonomous actors.

If there is a lesson from this event, it is the realisation that digital security is no longer separate from global politics.

This incident brings together technology, strategy, and responsibility at a single point. Thus begins a new phase of international relations, in which the question of who controls the algorithm is as important as who controls the territory.