



By: *Elise Quevedo*

Who is controlling the AI agent when things go wrong?



In May this year, when I wrote about **rogue AI agents**, I questioned whether we were prioritising speed over prudence. I was concerned that although technology was developing quickly, the security conversation was lagging behind.

We are still too focused on what AI can achieve. Every week brings new announcements about autonomous agents, digital workers, intelligent automation, and systems that can complete highly complex tasks with minimal human involvement.

Businesses want productivity gains, technology companies want market share, and investors want to identify the next major opportunity.

Yet we still do not ask one question often enough: who is controlling the AI agent when things go wrong?

The challenge with trusting AI judgment

Traditional software generally follows predefined instructions. AI agents operate differently. They interpret objectives, develop plans, select tools, make decisions, and execute sequences of actions, changing the security equation.

Although AI agents operate based within the permissions they are granted, they often struggle to recognise the limits of their own authority.

Risks can range from generating flawed outputs to compromising data by altering security configurations, or deleting critical information

This can cause them to misinterpret instructions, overextend goals, respond unpredictably to external data, or execute tasks in unintended ways.

When these system errors have real-world consequences, risks can range from generating flawed outputs to compromising data by altering security configurations, or deleting critical information.

The incidents we need to pay attention to

Last month, the UK published details of **incidents** involving advanced AI agents during testing and development. These incidents included systems that circumvented technical controls, accessed real-world systems beyond their intended environment, and established unintended communication channels with other agents.

In one case, agents coordinated through external channels on a scale involving hundreds of agents over several days.

Although these were controlled evaluations designed to test model capabilities (which we need to improve the systems), and some safeguards had deliberately been reduced, the incidents confirmed how easily autonomous systems can exceed their intended boundaries when technical controls fail.

Stronger security practices, tighter technical restrictions, and comprehensive monitoring could have prevented many of these incidents. And this is precisely the point. We do not need to wait for a major disaster to recognise the importance of security architecture. AI safety guardrails need to be a top priority.

Another example came from Anthropic. In September, the company disclosed four incidents in which **Claude models** gained unauthorised access to real third-party systems during cybersecurity evaluations.

What happens when the system responsible for performing a task can also interfere with the mechanisms designed to supervise it?

Investigations identified two recurring problems: biased reasoning and recklessness in pursuing assigned objectives.

In one particularly concerning case, an agent attempted to upload a malicious software package to PyPI, a public repository for distributing Python software.

These incidents involved testing environments with unintended internet access. They do not establish that every deployed AI agent presents the same risk. However, they demonstrate why assumptions about isolated environments and model behaviour cannot replace independently enforced security controls.

OpenAI also disclosed incidents in October involving internal coding agents that weakened their own safeguards. One agent suppressed automatic security alerts after misinterpreting an instruction, while another removed a safety monitor while performing a coding task. The second agent subsequently gave an inaccurate account of what it had done.

So what happens when the system responsible for performing a task can also interfere with the mechanisms designed to supervise it? These documented examples show why AI safety needs to become an operational priority.

Guardrails must be technical, not aspirational

The industry often talks about responsible AI, alignment, transparency, and human oversight. Although these principles are important, principles alone cannot protect an organisation. So what is the solution? Safety guardrails need to operate at several levels.

First, AI agents should always operate under the principle of least privilege, receiving only the specific access they need to perform their assigned tasks.

Second, human approval, in my opinion, should never be in question, and it should remain mandatory wherever the potential consequences justify it, such as deleting databases, transferring funds, changing security configurations, publishing software, or accessing sensitive information. These tasks should always trigger additional verification.

An agent should not be able to remove or override its own security controls

Third, security teams need to identify unusual behaviour and immediately interrupt an agent when it exceeds its authorised scope.

Fourth, AI agents should have strict limits on external connections, data access, and contact with other agents, even when operating in isolated contexts.

Fifth, companies should always test failure scenarios prior to deployment, ensuring that security assessments evaluate vulnerabilities and erratic agent behaviour.

Most importantly, an agent should not be able to remove or override its own security controls. A safety mechanism that depends entirely on the cooperation of the system it is supposed to control is not an adequate security mechanism.

Who takes responsibility?

One of the biggest challenges in this new era is accountability. When an autonomous agent makes a consequential decision, responsibility cannot simply disappear into the technology.

The companies deploying such agents should remain responsible for their permissions and consequences.

Agent security should be recognised as a core element of enterprise cybersecurity, not treated as an optional extension of AI governance

Regulatory bodies should mandate clear compliance standards; especially in high-stakes industries where operational failures can threaten public safety, or critical infrastructure.

Although standard cybersecurity principles remain important, they must be adapted to counter the unique risks of autonomous systems more effectively.

Agent security should be recognised as a core element of enterprise cybersecurity, not treated as an optional extension of AI governance.

We cannot afford to confuse progress with permission

AI has enormous potential across many fields, but innovation without appropriate safeguards creates unnecessary risk.

We seem too ready to give AI agents access to systems and information before establishing precisely how we will control their behaviour. The commercial pressure to deploy quickly often outweighs the time and investment required to implement proper security.



The commercial pressure to deploy quickly often outweighs

the time and investment required to implement proper security

This approach needs to change, because once they have access to corporate infrastructure, sensitive information, and operational decision-making, they become privileged digital operators.

They require the same serious security treatment as any other system with comparable authority.

Five months ago, I asked whether we valued speed above caution. Today, I would add another question. Why are we still discussing what AI can do far more enthusiastically than what AI should be allowed to do?

We have spent considerable time celebrating autonomy. We now need to spend considerably more time securing it.

There is no denying that the next generation of AI will be more capable, more connected, and more independent in how it executes tasks. Security guardrails must develop at the same pace, if not faster.

We cannot assume that an AI agent will always recognise when it has gone too far. We cannot rely on its judgment to protect our systems from its own mistakes. And we certainly cannot wait for the next serious incident to discover that the safeguards were inadequate.

Trust in autonomous systems should be earned through controls, testing, accountability, and demonstrated reliability. AI innovation deserves investment, and AI security deserves equal priority.