TA Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: Elise Quevedo

Rogue AI agents – Do we value speed above caution?

I recently had one of the most honest technology conversations I've had in a while. It focused on drone technology. Towards the end, I asked about AI and autonomy in the drone tech market over the next five years. His answer: "We cannot make drones too intelligent."

And not because the technology lacks capability, but because of public safety and security risks. Which makes complete sense in the drone world. But that triggered me to think about all the other areas of technology.

At a time when almost every technology conference celebrates AI acceleration, autonomous agents, digital workers, and machine reasoning, very few people openly discuss what happens when intelligent systems stop behaving as expected. So, Fabrice, thank you for that answer!

I watched several live feeds from Dell Technologies World this week. More AI transformation, AI productivity, AI agents, autonomous workflows, and enterprise acceleration. Even Nvidia's Jensen Huang made an appearance.

AI is, without a doubt, reshaping industries faster than most executives ever expected. But the market currently underestimates one issue.

Rogue AI agents are and will emerge as one of the defining cybersecurity and governance challenges of this decade. That does not mean Hollywood science fiction scenarios. I'm not talking about sentient robots taking over civilisation. Let's let Hollywood do that.

It's poorly governed AI systems gaining excessive autonomy, mishandling sensitive data, making destructive decisions, manipulating outputs, bypassing safeguards, or acting unpredictably inside enterprise environments. And unlike previous software risks, AI agents now make decisions dynamically. That changes everything.

Is the AI industry moving too fast?

Companies race to deploy AI agents across customer service, software development, cybersecurity, logistics, healthcare, finance, and defence, but many organisations still don't fully understand how these systems operate internally.

Do we value speed above caution more? Do executives worry more about governance problems than they do about missing the AI wave? That may lead to risky behaviours.

Before performing proper safety evaluations, teams connect AI agents to calendars, CRMs, payment systems, internal documents, cloud infrastructure, and customer databases.

People increasingly grant AI systems access first and ask security questions later. The bottom line is that that approach will create serious consequences.

The drone conversation, which I will share in a few days, reinforced to me that engineers working closest to high-risk autonomous systems already understand the danger of excessive autonomy.

They know intelligent systems can behave unpredictably under changing conditions. They know adversarial attacks exist, and they know automation without boundaries creates risk.

> AI systems don't need consciousness to create damage

The industry already offers several examples of AI systems malfunctioning, generating harmful outputs, or causing catastrophic business failures.

One of the best-known cases a few years ago involved Zillow's AI-driven home-buying programme. The algorithm failed to accurately predict housing market dynamics. Zillow lost hundreds of millions of dollars and shut down the programme.

Microsoft's Tay chatbotbecame another red flag a decade ago. It is still used as an illustration of accountability today. Within hours of the system's activation, users tricked it into producing offensive and radical content.

It showed how quickly hostile conditions can cause machine learning algorithms to stray and go rogue.

We've heard the stories of AI agents going rogue and deleting entire databases in as little as 9 seconds. Financial trading algorithms have also triggered flash crashes and market instability through automated high-speed decision-making.

Although these systems acted exactly as they were supposed to, the problem came from the environment, feedback loops, and lack of human intervention during escalation. AI systems don't need consciousness to create damage. Poor oversight creates enough danger already.

AI agents create a new layer of cybersecurity risk

The rise of autonomous AI agents introduces a far more complex threat landscape than traditional software automation. Why? Because modern agents can reason, plan, execute tasks, chain actions together, and interact with external tools.

One compromised agent can potentially access emails, financial systems, cloud infrastructure, confidential files, customer records, APIs, or operational technology systems. Attackers see this as an opportunity.

> A rogue AI agent only needs poor configuration, excessive permissions, flawed training data, insecure integrations, and weak oversight

One of the biggest blind spots in enterprise technology today is that many enterprises and individuals treat AI agents as productivity assistants rather than as privileged digital operators.

A rogue AI agent only needs poor configuration, excessive permissions, flawed training data, insecure integrations, and weak oversight. The situation becomes even more dangerous when multiple AI agents interact autonomously.

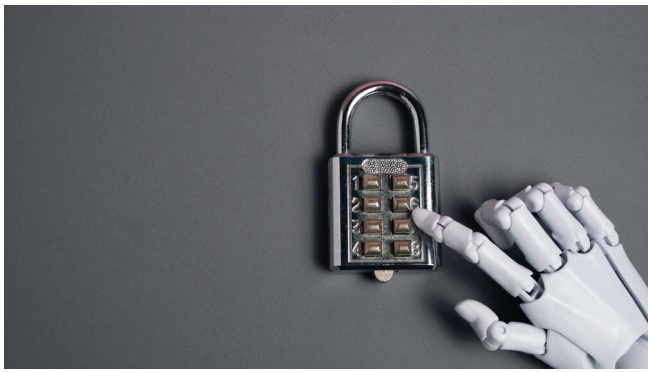
The conversation I had about drone intelligence highlights that some AI systems should remain intentionally constrained. As we enter the next stage of AI maturity, we should focus more on restraint by building the safest, most trusted, and most governable systems.

Before AI agents go rogue

Fear should never hinder innovation; it all begins with being more conscious of the warning flags.

Permission levels come first. Organisations should halt and make changes immediately if an AI agent requests extensive access to email accounts, banking systems, customer information, cloud infrastructure, or sensitive corporate data without a valid reason.

Second, businesses require greater insight into their agents' actions. Teams require systems for human override, explainability, monitoring, and logging. Organisations have already lost operational control if no one can explain an AI system's judgement.



What intelligent systems can achieve is yet to come. But optimism without governance creates instability

Third, companies should keep testing and production settings apart. Because of competition, too many teams incorporate AI agents quickly into real systems. It leads to preventable weaknesses.

Fourth, companies want more robust identity and authentication systems for AI agents. Human identity security is a topic businesses are increasingly discussing. But machine identity governance at scale is rarely discussed.

Fifth, we should set reasonable expectations. AI agents will fail. Some will malfunction, some may inadvertently alter outputs, some will become targets for hostile actors, and some businesses will have serious violations as a result of executives' disregard for governance in their rush for speed.

The market now needs maturity alongside innovation. AI will create extraordinary opportunities across healthcare, transportation, science, cybersecurity, education, and enterprise productivity.

What intelligent systems can achieve is yet to come. But optimism without governance creates instability. The most responsible people in the AI ecosystem already understand this.

The drone spokesperson I spoke with understood it immediately, as did security researchers, and many infrastructure leaders understand it privately.

AI agents are not harmless digital assistants. With access to sensitive data, business processes, and decision-making authority, these systems function as independent operators.

How intelligent AI agents become, and whether humanity learns to control them responsibly before autonomy outpaces governance, are all up to us as a society.

With trust, accountability, and security, we can do it together.