RSIS COMMENTARY

27 October 2025

Who's Accountable When AI Agents Go Rogue?

Asha Hemrajani and Ian Monteiro

RSiS S. RAJARATNAM SCHOOL OF INTERNATIONAL STUDIES
Nanyang Technological University, Singapore

Ponder the Improbable since 1996

RSIS S. RAJARATNAM SCHOOL OF INTERNATIONAL STUDIES
Nanyang Technological University, Singapore

RSIS COMMENTARY

www.rsis.edu.sg No. 212 – 27 October 2025

Who's Accountable When AI Agents Go Rogue?

By Asha Hemrajani and Ian Monteiro

SYNOPSIS

The rise of autonomous AI systems has revealed a new frontier in cybersecurity risk, expanding attack surfaces and blurring accountability. Safe and responsible deployment has hence become a defining cybersecurity challenge. Governance of non-human identities and adaptive, policy-driven controls to detect and contain attacks on AI models, apps, and workflows will be needed to establish trusted autonomy.

COMMENTARY

Earlier this year, security researchers proved that an artificial intelligence (AI) assistant could be hijacked through something as ordinary as a calendar invite. Hidden within the invitation was a set of malicious instructions that, once triggered, caused connected lights to flicker, shutters to open, and files to be accessed without the user's consent.

What began as a controlled experiment quickly revealed a new frontier in cybersecurity risk, where AI systems are not just tools for attackers but potential targets in their own right. As AI becomes more autonomous, able to plan and act across digital and physical environments, the implications for security will be far-reaching.

The line between human and machine agency is blurring, and the time needed to exploit vulnerabilities is shrinking. For businesses and governments, this signals a fundamental change in how digital risk must be managed.

This shift from passive tools to autonomous agents is ongoing. Agentic systems are already deployed in banking, e-commerce and logistics to streamline operations, detect fraud and make real-time decisions.

As these agents interact with enterprise systems, other agents and humans, the cybersecurity attack surface expands. Malicious agents can exploit the same interfaces as legitimate ones, using new threats such as impersonation attacks, prompt injections and data exfiltration (theft). Safeguarding agentic AI in enterprise systems is therefore emerging as a defining cybersecurity challenge.

Cybersecurity as Strategic Enabler

Governments and enterprises are now seeking ways to capture the benefits of AI innovation while managing the growing spectrum of risk it creates. The discussion is increasingly on how to deploy it securely and responsibly.

Traditional cybersecurity frameworks were designed for systems with predictable behaviours. Agentic AI breaks that predictability. It learns, adapts and operates with varying degrees of autonomy, creating new layers of uncertainty that static defences cannot contain.

For governments and large enterprises operating critical infrastructure, this shift requires a fundamental change in mindset. As agentic AI becomes embedded in decision-making, operations and citizen services, cybersecurity must evolve from a defensive function to a strategic enabler of trusted autonomy.

This demands a shift to adaptive, context-aware security with clear human oversight and escalation management, moving beyond static defences to maintain the trustworthiness of systems that influence decisions at a national scale.

Foundational concepts in cybersecurity, such as identity, data, and attack surfaces, are taking on new and evolving dimensions. Even established frameworks like “zero trust” are being re-examined as the rise of AI exposes contradictions that demand rethinking and adaptation.

Reframing Digital Risk Governance

Indeed, governance frameworks must evolve alongside technology. Two issues are becoming urgent.

First, the spectrum of autonomy must be understood. Agentic behaviour is not a binary state. Treating a basic automation script as equivalent to a self-directing system results in misplaced controls and uneven risk management. Oversight and safeguards should correspond to degrees of autonomy, not broad labels.

Second, accountability must be redefined. If an agentic AI system executes an action that is harmful, who should bear responsibility? Without clear boundaries, legal and ethical gaps will persist, and adversaries may exploit them. Boards, chief information security officers and regulators need shared accountability models that reflect how agentic AI systems work.

These questions are already visible in data governance disputes, algorithmic bias cases, and AI incidents where AI systems have behaved in unexpected ways. Unless accountability frameworks get better defined, accountability gaps will widen.

Securing Agentic AI in Critical Infrastructure

Agentic AI deployment in critical infrastructure entities raises unique risks. These systems promise gains in efficiency and resilience, but their vulnerabilities could cause cascading disruptions if compromised. Protecting them requires new approaches to securing AI apps and agents. It is therefore essential that critical infrastructure entities retain control as they adopt more autonomous AI-driven systems.

The focus must then be on detecting and stopping attacks on AI models, apps, and agentic-AI workflows. Policy controls for AI use, including blocking risky requests, preventing data leaks in apps and detecting unsanctioned AI agents, are also essential.

Equally important is ensuring resilience by governing the non-human identities (NHIs), the digital identities backbone of agentic AI. Enterprises will need to exercise proper oversight of NHIs through access control, guardrails and traceability.

Convening for Resilience in Agentic AI

Trust will not be built by algorithms alone; technology is only as trustworthy as the intent and integrity of the people who create and govern it. The rise of agentic AI exposes the limitations of current frameworks and demands new approaches grounded in foresight, accountability and collaboration. Businesses that recognise this shift will be better protected and positioned to lead in the next chapter of digital transformation.

Asha Hemrajani is a Senior Fellow at the S. Rajaratnam School of International Studies (RSIS), Nanyang Technological University (NTU). Ian Monteiro is the Chief Executive Officer and founder of Image Engine, organiser of the GovWare Conference and Exhibition 2025. *This commentary was originally published by The Business Times on 21 October 2025. It is republished here with permission.*

S. Rajaratnam School of International Studies, NTU Singapore
Block S4, Level B3, 50 Nanyang Avenue, Singapore 639798

Please share this publication with your friends. They can subscribe to RSIS publications by scanning the QR Code below.