Why zero trust must extend to AI agents in the autonomous enterprise

As agentic AI scales across enterprises, Birlasoft argues that Zero Trust, human oversight, and machine-speed governance are critical to securing autonomy at scale.

By Shrikanth G

As enterprises race to operationalise agentic AI, automation is no longer confined to efficiency gains or isolated use cases. Autonomous agents are beginning to operate inside security operations, identity systems, and enterprise workflows, observing signals, making decisions, and acting at machine speed. This shift promises scale and resilience, but it also introduces new risks that traditional governance and security models were never designed to handle.

Ganesan Karuppanaicker, Chief Technology Officer at Birlasoft, argues that the real challenge of agentic AI lies not in capability, but in control. As multi-agent systems become embedded across enterprise environments, organisations must rethink security, trust, and oversight, extending Zero Trust principles beyond people and devices to AI systems themselves.

FROM ASSISTANTS TO AUTONOMOUS DIGITAL TEAMMATES

Agentic AI is transforming enterprise automation at unprecedented speed. Organisations are moving rapidly from traditional AI assistants to autonomous digital teammates, systems capable of observing, deciding, acting, and continuously learning within core business processes. This shift is accelerating across Indian enterprises, with Deloitte noting that more than 80% of organisations are actively exploring autonomous agents as a strategic pillar of their generative AI roadmap.

These systems promise continuous decision-making and faster outcomes, but autonomy at scale demands a fundamentally different approach to risk, governance, and security.

GANESAN KARUPPANAICKER
Chief Technology Officer, Birlasoft

SECURITY GUARDRAILS FOR AUTONOMOUS AGENTS

Autonomous security agents require strict guardrails to avoid inadvertently opening new vulnerabilities. Applying Zero Trust to AI is critical, where each agent is given minimal access, subjected to continuous checks, and required to prove itself at every step. Organisations must adopt human-in-the-loop models wherever high-stakes actions or risky operations are involved.

> Red teaming, staged scenarios, and dark mode trials help expose failure paths before agents are deployed at scale.

Rigorous testing is equally important. Red teaming, staged scenarios, and dark mode trials help expose failure paths before agents are deployed at scale. AI governance must operate at machine speed, allowing human experts to focus on strategy, threat hunting, and validation of AI decisions. With guardrails, restricted permissions, and human oversight in place, organisations can confidently leverage autonomous agents.

SECURING MULTI-AGENT ECOSYSTEMS

Multi-agent ecosystems introduce new machine-to-machine attack surfaces. Secure orchestration, role-based isolation, adversarial testing, and continuous oversight are essential to prevent misuse or hijacking. In multi-agent architectures, every agent interaction becomes a potential attack vector.

By treating each agent as untrusted by default and forcing it to continually prove its integrity, while layering defensive AI to monitor for tampering, organisations can benefit from multi-agent collaboration without triggering a domino-effect breach.

ZERO TRUST FOR ALGORITHMIC ACTORS

In this new paradigm, Zero Trust must extend beyond people and devices to include AI systems themselves. Agents must authenticate their identity, understand contextual intent, and clearly justify planned actions before execution. There can be no implicit trust for algorithmic actors. Each agent must be treated like a privileged user, operating in a trust-nothing, monitor-everything mode.

A HUMAN-LED, AI-AMPLIFIED SECURITY FUTURE

The impact of agentic AI is especially significant in cybersecurity. Well-designed agentic workflows are already delivering measurable gains by improving alert precision, reducing response times, enhancing identity analytics, and enabling continuous control monitoring. By taking over repeatable, time-sensitive tasks, agents free security operations centre teams to focus on adaptive oversight, higher-order judgement, and complex threat scenarios.

Agentic AI's future in cybersecurity is not about replacing human expertise, but amplifying it. Smart agents will manage routine responses autonomously, while analysts retain situational control, refine decision logic, and drive strategic improvements. Together, this creates a security model that is faster, smarter, safer, and ready for the autonomous era ahead.

shrikanthg@cybermedia.co.in