

The reset button
is lying to you
p7

Why smart homes
need graceful
degradation p30

The EV battery
rules owners
get wrong p48

The next SaaS giant
may own the AI
operating layer p64

₹125

VOL. 39 | ISSUE 9 | SEPTEMBER 2026

www.pcquest.com

PCQUEST

TECH TODAY & TOMORROW

 CyberMedia



THE REAL TEST OF SMART

Smart tech is everywhere.
The real test is whether it
reduces work, risk, and friction.



76 pages including cover



Special Subscription offer on page 56

AI agents need security before autonomy

Tarun Wig

pcquest@cybermedia.co.in



AI agents can act at machine speed, and so can their mistakes. As enterprises hand AI more control, security guardrails, access limits, human oversight, and digital kill switches become critical to safe automation.

The enterprise artificial intelligence (AI) ecosystem is constantly evolving, shifting away from the world of passive, conversational generative AI (GenAI) chatbots that usually summarize documents or draft emails. Now, the industry is quickly entering an age of autonomous AI agents, which are intelligent systems that can reason, plan, and execute complex workflows without constant human involvement. AI agents provide substantial operational efficiency improvements by independently managing the IT service desk, optimizing supply chains, and executing automated threat hunting in the cybersecurity space.

However, the shift from content generation to autonomous execution creates new and significant risks. As an example, if an AI system moves forward from writing a report to automatically updating an enterprise database or configuring a firewall, any errors made by the AI system potentially scale exponentially. Indian enterprises do not have any other option than to build strong security guardrails. It is not an afterthought anymore. It has become a prerequisite to protect the integrity of the business.

▼ The anatomy of the agentic security gap

Traditional cybersecurity paradigms are built around securing human actions and predictable software logic. Autonomous AI agents disrupt this entirely by introducing a dynamic, unpredictable execution layer into daily business operations. This changes the enterprise risk environment in three important ways:

▼ Power of autonomy

AI agents differ from traditional software because they make their own decisions about how to reach a goal, rather than following preset rules. For example, if an unguided agent is asked to optimize server performance, it may interpret that request to mean that deleting old log files to free up storage will help with its task and will do so even if this means violating company policy.

▼ Risk of goal divergence

Agents operate via complex reasoning



TARUN WIG,
Co-founder and CEO, Innefu Labs

chains. If malicious actors manipulate the inputs (via advanced prompt injection attacks), they can hijack the agent's logic. A hijacked agent can be easily manipulated into the exfiltration of proprietary data, unauthorized financial transactions, or system configuration sabotage.

▼ Cascading machine-speed failures

When systems run at machine speeds at massive scales, a single logic failure or a security breach can have a cascading impact across multiple enterprise systems. This all may occur even before there is human intervention to detect an anomaly.

▼ Enforcing the guardrails: A blueprint for secure scale

To safely use the power of agentic AI, enterprises must wrap these autonomous systems in a continuous, multilayered security approach.

1. Identity and access boundaries

Treat AI agents exactly like privileged human employees. Every agent requires a unique, verifiable machine identity. Agents must operate on short-lived, task-specific tokens that automatically expire upon completion of the workload.

2. Multitiered action authorization

High-impact actions, such as authorizing financial payouts, changing user permissions, or modifying production code, should require human approval or active human oversight.

3. The principle of least agency

An agent should only possess the tools absolutely necessary for its specific job. If an agent's task is to analyze customer feedback trends, its application programming interface (API) permissions must be strictly isolated from modifying client accounts or viewing billing databases. Minimizing the blast radius is essential.

4. Real-time logic and behavioral filtering

Deploy real-time monitoring solutions that scan both the inputs feeding into the agent and the internal reasoning steps the agent takes. This allows organizations to catch adversarial prompt injections or alignment drift before they translate into backend system actions.

5. Hardcoded autonomy limits

Organizations need to define clearly what operational thresholds will apply to agents. AI agents should have clear boundaries indicating when they have exceeded their authority. If a request is ambiguous, exceeds the agent's authority, or involves a high-risk action, the agent should decline to act and refer the request to a qualified human operator.

6. Context-aware data governance

Organizations should use clear data-governance controls, including masking and redaction, to prevent agents from exposing personally identifiable information (PII), trade secrets, and other sensitive data.

7. Automated runtime circuit breakers

Just like an electrical grid uses circuit breakers to prevent fires, agentic ecosystems need digital kill switches. If an agent begins looping destructively or executing unauthorized API calls, automated filters must instantly freeze the agent and revert the environment to a safe state.

▼ Conclusion

Securing AI agents is not a one-time configuration, but rather a continuous governance framework that uses modern compliance benchmarks such as the NIST AI Risk Management Framework. It is necessary for organizations to have a risk management framework (RMF) in place. Additionally, organizations are required to maintain a living AI agent inventory and an AI Bill of Materials (AIBOM) to identify exactly what models, data sources, and tools are interacting within their networks.

The future of businesses is dependent upon those organizations that can automate safely. By implementing strong security guardrails into our AI architectures, Indian organizations will confidently advance from experimental AI pilots to resilient, autonomous, and secure digital powerhouses. ■

The author is Co-founder and CEO, Innefu Labs

