Tuesday, September 15, 2026 | 12:23 PM IST

Businesses put up guard rails in deploying AI, prioritise governance

Fear of rogue agents is prompting firms to limit autonomy and prioritise governance

Imaging: Ajaya Mohanty

Avik Das |Aashish Aryan |

6 min read Last Updated : Sep 13 2026 | 9:55 PM IST

An exceptionally smart agent goes rogue and operates outside its organisational mandate. Depending on your interests, that describes either the plot of the James Bond movies Skyfall and Spectre or PHASEONE10841, OpenAI's rogue agent that in July escaped its controlled environment and hacked Hugging Face, another artificial intelligence company.

Enterprises deploying AI agents are working on the assumption that these powerful tools could stray from their mandated behaviour. Rather than handing them complete autonomy, companies are restricting what agents can do, what systems they can access, and when humans need to intervene.

Adoption of agentic AI — systems that can independently plan, execute and adapt multi-step tasks without a human prompt — is accelerating, but autonomy remains constrained by how work is structured, governed and trusted, said a report by Infosys, the Indian information technology (IT) services company, and HfS Research, a global research and advisory firm, earlier this year.

Safety first

"Enterprises are choosing reliability over reach, and the data makes that choice visible. Almost 60 per cent say their most advanced agents are performing simple, rules-based tasks, and only 16 per cent report enterprise-wide deployment. Agents are present, but they are largely confined to safe, scripted terrain," said the report.

Also Read

AI helps students with work, but more use doesn't mean better scores: Study

Beijing hits back at Anthropic CEO's call to curb China's AI development

COAI, British High Commission sign pact on AI, digital trust, connectivity

ICBC, HSBC among ma ByteDance's $30 billio

The report highlighted that enterprises are taking this approach due to a lack of governance, data access, and accountability.

The report, which took insights from 500 global enterprises, noted that another reason for the cautious adoption of agentic AI is cybersecurity. "Only 16 per cent of enterprises say their cybersecurity infrastructure could readily handle agent-specific threats, and just 12 per cent are comfortable giving agents broad access to sensitive data — evidence that containment, not capability, is the current ceiling."

Babak Hodjat, chief AI officer at IT firm Cognizant, warned that autonomous agents and systems interacting with each other and their environment could trigger unintended and risky behaviour.

Despite advances in AI models and agents, enterprises currently lack justification to adopt the most powerful ones, said Arun Ramchandran, chief executive officer (CEO) of digital product engineering platform QBurst.

He believes most enterprises should use lower-capability, less expensive AI models and agents for now, as the issue with frontier, high-capability options is not just security.

"Security and guard rails are important but there is an inherent conflict between the alignment and capability of these models. By that, we mean rewards and reinforcement learning. This is an alignment issue at the heart of how the model was created," he said.

The debate over AI risks intensified on September 9, when former Anthropic employee Jacob Coxon alleged after resigning that neither Anthropic nor his previous employer, OpenAI, was developing AI responsibly.

Writing on X (formerly Twitter), Coxon alleged that both companies were "racing straight to self-improving superintelligence and gambling" with the lives of ordinary people.

Coxon's post racked up over 164 million views, sparking a debate. Supporters backed his concerns and demanded urgent government regulation, while critics dismissed the fears, arguing AI agents remain under human control and can be shut down.

For enterprises, the immediate concern is less about an AI agent developing its own agenda and more about the risks of a probabilistic system making unintended decisions with access to critical business operations.

Enterprises are debating whether to deploy AI agents across their operations, said Vineet Moroney, chief transformation officer of Xoriant, a software and technology firm.

Deploying AI agents in deterministic processes — where a specific input yields only one output — keeps error rates contained. In probabilistic systems such as large language models (LLMs), where single inputs yield varied outputs, the error rates for AI agents increase, though so does a company's ability to refine and tailor the process, Moroney said.

Calvin C Newport, a professor of computer science at Georgetown University, argued in a recent blog post that AI systems that operate by autonomously executing LLM-generated plans are a "really bad idea".

"Not because these systems are devious, or malicious, or inventing their own agendas, but because LLM output is unpredictable and non-normative," Newport said.

The question for enterprises, therefore, is increasingly about how much autonomy an AI agent should have and how to contain the damage if it makes a wrong decision.

While some experts argue for a "kill switch", Sharda Tickoo, country manager for India and SAARC at TrendAI, a computer and network security company, and others believe it should be the last line of defence for enterprises.

Autonomy debate

"I think the bigger challenge is designing agents where we are able to contain the blast radius if a wrong decision is made. Given how the whole agentic AI architecture is built, there is a discussion about having one super agent versus specialised agents doing specific jobs. Essentially, you are narrowing the scope of a single agent rather than giving one super-agent full access," Tickoo said. Once an LLM is connected to an enterprise's tools, application programming interfaces, data and transaction systems, the system architecture has to assume that the model and the agent trained on the model will occasionally stray from ideal behaviour, said Amit Verma, founding head of technology at Neuron7.ai.

"Different actions can have different autonomy levels: Some can execute directly, some require confirmation, and some should never be available to the model at all. I suspect this becomes particularly important in regulated environments where auditability and reproducibility matter as much as model accuracy," he said.

"I believe the key controls are around identity, in terms of what is permissible. I also believe the whole debate around whether agent architecture should have a kill switch reflects the fact that we should never underestimate the risk of it going rogue, even if you put in some guardrails," Tickoo said.

AI models will improve but there is already significant scope to build businesses using existing capabilities, said Krish Ramineni, CEO of Fireflies.ai.

"Even if the models did not improve for more than five years from today, there are still going to be incredible businesses that will be built on top of these existing models," Ramineni told Business Standard in a recent interview.

Small firms may find it harder to tackle rogue AI agents, he said.

For enterprises, therefore, the question may ultimately be less about whether an AI agent can go rogue and more about designing systems on the assumption that, at some point, one will.

More From This Section

BRICS shared satellite system may cut space debris, optimise costs: Experts

Apple iPhone 17 sees strong demand as buyers rush to beat steep price rise

Anthropic CEO Amodei sounds alarm, urges slowdown in AI development

Another Anthropic res AI race risks after quit

Topics : Artificial intelligence | Technology | governance

Buzzing : Tata Sons Listing | Delhi Traffic Advisory | Weather Update | Bank Strike Today | Top 10 Chatgpt 80s Look Prompt | NSE IPO Price Band | iPhone 18 Pro series debuts | Br

Don't miss the most important news and views of the day. Get them on our Telegram channel

First Published: Sep 13 2026 | 9:55 PM IST

Read Comments (0) Add Comment

Next Story

Home / Opinion / Columns / Smart tech powers India's energy needs

Smart tech powers India's energy needs

Smart grid technology and modern transmission infra are critical for country's renewable-powered economic agenda