UNDER PRESSURE

Mohammad Ismail *asks is Agentic AI in arrested development?*

Talk of an AI bubble has dominated the headlines over recent months. Concerns are growing that the level of investment being pumped into AI companies is out of kilter and that, much like we saw in the dot. com era, this could lead to a crash. Yet others are arguing the exact opposite. They claim the sector is continuing to see exponential growth, with the technology evolving so fast that it doesn't even bear comparison with previous technological milestones such as the internet and cloud, more than justifying the continued influx of investment.

According to the Cloud Security Alliance (CSA), AI is undergoing significant evolutionary cycles every 3-6 months, introducing new capabilities and reshaping expectations. The problem is that this constant shapeshifting is also making it extremely hard for organisations to keep up. Projects are already struggling, with an MIT report finding that 95 percent of generative AI projects are failing to deliver business value and its successor, agentic AI which will see AI given far greater autonomy, is also facing a tough time. Gartner has predicted that 40 percent of these agentic AI projects will be cancelled by 2027 with these projects proving too difficult to get off the ground.

So why are AI projects foundering? As with most IT development, agentic AI requires careful planning with respect to timescales, resource and risk. The problem is that AI is a moving feast, resulting in additional costs such as upskilling developers or retooling agentic solutions as specifications change and there's no real blueprint to follow. Consequently, the most commonly cited reasons for failed agentic AI projects range from unclear or unrealistic business objectives to governance, compliance, and security risks, and workflow integration challenges.

A Chinese state-sponsored group used AI's 'agentic' capabilities to execute cyberattacks

It's a tough hill to climb because the business is under pressure to push these projects into production as soon as possible to achieve productivity gains and revenue growth. Gartner has also predicted that 15 percent or more of day-to-day workplace decisions will be made autonomously by agentic AI systems by 2028, revealing that the technology offers real advantages. But scaling a prototype is no small feat and many are finding that cracks can soon start to appear. Access and authentication is a prime example as all too often agents are being given access to core systems without clear boundaries or context, which leaves the proverbial door wide open to data leakage, misuse and compliance issues.

Stories of AI abuse are also now becoming more common, both in the form of known attack types and novel forms. Over the course of the past six months, we've seen the emergence of AI-driven polymorphic ransomware (ie: ransomware that can rewrite itself, mutate its behaviour and evade detection by signature-based systems). One example is PromptLock, detected in August 2025, which uses a prompt-injection attack sent via the Ollama open source API to trick Large Language Models (LLMs) into assisting in the ransomware attack.

Similarly, the Google Threat Intelligence Group (GITG) reported in November that malware families are now being used to craft 'just in time' malware, with PromptFlux and PromptSteal both using LLMs to create malicious functions on demand rather than hard-coding them into the malware. And Anthropic's threat intelligence team also announced during the same month that they had detected the first case of an AI-orchestrated cyber espionage campaign that used the Claude Code tool to target approximately 30 organisations globally.

Such attacks are running rings around traditional security tools, which are simply not designed to detect or handle them – particularly as the infrastructure itself is now changing. Agentic AI uses Modern Context Protocol (MCP) servers that act as a conduit for AI agents, allowing them to easily access information and services from disparate sources, but since the release of the MCP standard in November 2024, the focus has primarily been on how it can open the floodgates to those resources, not on how they can be secured.

Any business looking to use AI now needs to come to grips with these mechanisms and securing them. Typically, the organisation will either create its own MCP servers and/or connect to third-party versions, but they don't know if those servers are secure. When developers use those servers, they can create backdoors into enterprise systems, which enable 'typosquatting' and other similar attacks that impersonate legitimate integrations and can exfiltrate data or cause other damage while appearing to function normally. We saw this recently in the form of a malicious MCP server used to steal emails in a rogue 'postmark-mcp' package in npm, which copied an official Postmark Labs library of the same name.

MCP servers effectively centralise access to multiple sensitive services such as email, databases and cloud systems, which means that attackers who compromise a single server can gain access across the enterprise. This makes stored OAuth tokens high-value targets, so protecting them is a must which means that rather than agentic AI projects sidestepping zero trust, they should be included and subjected to the same continuous verification of identity and permissions. If agent permissions are not nailed down, AI can pull sensitive data across service boundaries. The organisation can also lose visibility into what data agents touch and where it goes – risking lost intellectual property and customer data, compliance violations, and more.

THE GATEWAY PROVIDES A HUB VIA WHICH TO MONITOR EVERY AI AGENT INTERACTION

For these reasons, it's crucial to have a trusted registry of MCP servers that have been vetted and are safe to connect to. MCP server creation also needs to be governed by server development and usage policies. And the business needs to allow for the standard itself to change. MCP can and will evolve, so from a development perspective the business will need to be able to perform updates to ensure applications remain compatible without the need to re-engineer them.

Other idiosyncrasies associated with AI agents should be considered too, such as the potential for business logic abuse. As we saw with APIs, attackers are quick to learn and exploit M2M communications, but with AI agents the scope for abuse is much higher because they are so adaptable. This means that AI agents can be used to mimic legitimate users in order to bypass defences that lack business context. But it doesn't stop there. Because they're AI-driven, they're able to identify and exploit logic flaws faster than human attackers and can prioritise profitable abuse paths, which means that the organisation will need to fight fire with fire to combat such abuses. That will require the real-time monitoring user-agent-AI traffic to detect abuse.

Given these issues, it's no wonder that organisations are struggling to get their projects off the ground. But it is possible to make it easier to create, deploy, and manage AI by centralising the process. AI gateways, which were initially designed to act as a conduit for AI traffic, are now becoming much more geared towards development and governance and can be used to make any application agent-ready or any endpoint MCP-compatible without the need for coding.

When it comes to MCP, a gateway can be used to both spin up or connect to servers and by comparing third-party servers against a trusted list of vetted servers and APIs, it can prevent connection to bogus servers. Furthermore, as the gateway is abstracted from each of the components in the chain, it can be used to manage API updates or changes to MCP or any other protocols for that matter, ensuring applications always remain compatible.

Using a gateway also provides the ability to integrate with OAuth-compliant identity providers,

so can be used to enforce identity-based access to systems and data, preventing unauthorised AI access. And it can align the AI with zero trust network architectures to ensure continuous validation, avoiding the problem of AI projects blowing a hole in zero trust protection.

ANY BUSINESS LOOKING TO USE AI NEEDS TO COME TO GRIPS WITH SECURING RESOURCES

Finally, the gateway provides a hub via which to monitor every AI agent interaction. By logging interactions between user, API and agent, it's possible to see which applications are being access by agents, what API calls they make and what data they touch. This then allows suspicious activity to be flagged, investigated and mitigated before a data leak or attack can occur.

Gateways are therefore likely to become a key lynchpin in the development of AI and may well become the determining factor in whether projects succeed or fail because they effectively allow the organisation to oversee and control those variables of time, resource and risk. But it's not simply a matter of selecting a solution. It's such a volatile market and AI-washing is so rife that organisations need to be mindful of other factors. These include seeking provider/s that can offer assurances in the form of experience in the market such as from an application or API background, that offer bulletproof SLAs and support, and have a long-term roadmap in place with respect to the expected evolution of the market.

Looking to the future, what is clear is that businesses cannot afford to continue on their current trajectory, ploughing time and money into AI. Project development lifecycles of six to 18 months are simply not sustainable in such a fast-paced market, rendering projects obsolete before they've even left the starting blocks. In fact, it is estimated that AI is four times faster than Moore's Law, which sees computing power double every four years, giving some indication of just how fast-paced technological development is. But nor can the business afford to sit on its hands, while their competitors steal a march.

The upshot of both these arguments is that businesses need to prioritise security in project development from the get-go. In doing so they may just find a way to circumvent many of the obstacles that have been preventing them from achieving a workable AI solution and finally be able to capitalise on promised productivity gains and revenue growth ●

Mohammad Ismail is VP of EMEA at Cequence Security and has extensive cybersecurity experience, particularly in the Identity and Access Management space.

AI agents can be used to mimic legitimate users in order to bypass defences

Picture credit: AdobeStock