LUMO INSIGHTS

Tuesday, 16 June 2026

AI agents evolve: from novelty to necessary business tools

Executive Summary

In the past two days, AI agents have taken big strides towards enterprise-grade maturity. From surprising new tools that control AI's cost and credentials, to emergent strategies for mixing models and standardising workflows, the focus is shifting from hype to practical integration. The latest developments signal that businesses can now address many of the hidden obstacles holding back AI automation, making autonomous workflows more secure, cost-effective, and scalable.

Securing Trust in autonomous workflows

For organisations to let AI agents handle sensitive tasks, they must first solve a fundamental issue: trust. Empowering an autonomous agent to perform meaningful work often means granting it access to private data, corporate systems, or customer information. Historically, this created a Catch-22. The more capability and autonomy an AI agent had, the more you had to expose your digital keys - API tokens, passwords, customer records - and therefore the greater the risk of leaks or misuse. Many companies have dealt with this by hard-coding credentials in code or giving agents broad access, strategies that are both insecure and hard to audit.

This week brought a significant step towards eliminating that roadblock. 1Password, whose password vault is already trusted by over 180,000 businesses (www.morningstar.com [1]), announced a new Credential Broker service that securely injects credentials into automated tasks and AI tools when they're needed (www.morningstar.com [2]). Instead of a developer or data scientist copying an API key into a prompt or script, the agent can request a short-lived secret from the 1Password vault at runtime via the broker. According to 1Password’s CTO, Nancy Wang, the goal is to 'close the gap between where credentials are protected and where access happens' by moving away from secrets scattered in code and towards an identity-driven, auditable delivery of credentials (www.helpnetsecurity.com [3]). This approach reduces the attack surface for breaches and makes it far more feasible to let AI agents perform tasks like querying databases or third-party services without putting enterprise data at risk.

We're also seeing legacy enterprise platforms adapt to this new agentic era with a focus on governance. On Sunday, Oracle published a technical deep dive into its Oracle AI Database Private Agent Factory, illustrating how multiple AI agents can be orchestrated to work directly with corporate data and tools under database-grade security controls (www.aitechblog.blog [4]). Oracle's approach keeps AI close to the data - running on the company's own cloud or on-prem systems - using an open Model Context Protocol to integrate external AI models without exposing sensitive information beyond the firewall. The message is clear: as AI agents move from novelty to real production use, enterprises are demanding (and now getting) solutions that let them maintain control over credentials, data, and

compliance.

References:

[1] www.morningstar.com — https://www.morningstar.com/news/business-wire/20260615920466/1password-introduces-credential-broker-building-a-secure-credentialing-layer-for-humans-machines-and-ai-agents#:~:text=Building%20on%20the%201Password%20vault,machine%20workloads%2C%20and%20AI%20agents
[2] www.morningstar.com — https://www.morningstar.com/news/business-wire/20260615920466/1password-introduces-credential-broker-building-a-secure-credentialing-layer-for-humans-machines-and-ai-agents#:~:text=Building%20on%20the%201Password%20vault,machine%20workloads%2C%20and%20AI%20agents
[3] www.helpnetsecurity.com — https://www.helpnetsecurity.com/2026/06/15/1password-credential-broker-reduces-secret-sprawl-through-identity-based-credential-delivery/#:~:text=needs%20to%20happen,secrets%20to%20brokered%20credentials%20The
[4] www.aitechblog.blog — https://www.aitechblog.blog/post/oracle-ai-database-private-agent-factory-turning-enterprise-data-into-secure-ai-agents#:~:text=available%20on%20Oracle%20Cloud%20Marketplace%2C,responses%20Secure%20access%20Deploy%20at

Taming the costs of autonomous AI

Another surprise for many leaders has been the hidden cost of enthusiastic AI automation. AI agents don't operate on a fixed salary - they consume computing resources and API calls, sometimes voraciously. Some companies have discovered this the hard way, watching their 'token budgets rise faster than payroll' when they unleash always-on AI processes (intellyx.com [1]). Recognising this, Amazon Web Services has introduced new tools to bring financial discipline to AI agents.

At its New York City summit this week, AWS unveiled an AI-powered FinOps Agent, now in preview, designed to monitor and control cloud spending by autonomous workflows (aws.amazon.com [2]). This agent can answer natural-language questions about AWS costs and generate detailed cost reports for finance and engineering teams. Beyond reporting, it actively flags waste (like underutilised resources) and can even open tickets in Jira to kick-start cost-saving changes (aws.amazon.com [3]). When a spending anomaly is detected, the FinOps Agent will automatically investigate the root cause and post its findings to a Slack channel, enabling teams to respond immediately without manual triage (aws.amazon.com [4]). By embedding cost awareness into automation itself, AWS is giving enterprises a new way to catch runaway AI expenses in real time - before they blow the budget.

Keeping AI on budget goes hand-in-hand with keeping it under observation. AWS also rolled out a novel integration for its OpenSearch monitoring service that embeds MCP (Model Context Protocol) "Apps" - specialised observability tools an AI agent can call for logs, metrics, and traces during incidents (aws.amazon.com [5]) (aws.amazon.com [6]). This means a troubleshooting agent running in a developer's environment (like an IDE) could automatically pull relevant log data or performance charts to diagnose a problem, offering a glimpse of how AI might soon assist in DevOps and IT operations. The bigger picture: cloud providers are starting to treat autonomous agents as first-class citizens of the enterprise tech stack, complete with their own cost centres and monitoring needs. Forward-looking leaders should ensure their AI initiatives include similar guardrails and visibility - so the benefits of automation aren't undermined by budget surprises or hidden failures.

References:

[1] intellyx.com — https://intellyx.com/2026/06/14/bloomfilter-process-and-context-optimization-for-ai-and-agentic-development-adoption/#:~:text=can%20drag%20down%20long,as%20making%20complex%20decisions%20or
[2] aws.amazon.com — https://aws.amazon.com/about-aws/whats-new/2026/06/aws-finops-agent-preview/#:~:text=available%20in%20preview%20Posted%20on%3A,and%20runs%20recurring%20FinOps%20workflows
[3] aws.amazon.com — https://aws.amazon.com/about-aws/whats-new/2026/06/aws-finops-agent-preview/#:~:text=teams,Virginia%29%20Region%20and
[4] aws.amazon.com — https://aws.amazon.com/about-aws/whats-new/2026/06/aws-finops-agent-preview/#:~:text=teams,Virginia%29%20Region%20and
[5] aws.amazon.com — https://aws.amazon.com/about-aws/whats-new/2026/06/opensearch-agentic-observability-mcp-app/#:~:text=Amazon%20OpenSearch%20Service%20launches%20MCP,tool%20call%20returns%20a%20dual
[6] aws.amazon.com — https://aws.amazon.com/about-aws/whats-new/2026/06/opensearch-agentic-observability-mcp-app/#:~:text=response%2C%20a%20concise%20text%20summary,metrics%20and%20trace%20investigation%2C%20service

Mixing models to boost efficiency

Generative AI keeps getting more capable - but with greater capability often comes greater cost. The latest top-tier models from the likes of OpenAI, Google, or Anthropic are extraordinarily powerful, but also expensive and sometimes overkill for simpler tasks. This week, however, brought evidence of a more nuanced strategy: using a combination of AI models so that each task is handled by the most appropriate engine, rather than throwing the biggest model at every problem. As the Wall Street Journal has reported, businesses are increasingly routing work across 'a mix of open-weight and frontier commercial models,' saving the most powerful (and costly) systems for only the toughest challenges (cast.ai [1]).

Cloud automation firm Cast AI provided a timely example of this approach with its "Kimchi" coding agent. On Monday, Cast AI announced that Kimchi is the first autonomous coding system to integrate a new open-source model called MiniMax M3 - one of the most advanced publicly available AI models for code and long-context tasks (cast.ai [2]). MiniMax M3 can handle up to one million tokens of context (essentially reading an entire codebase or lengthy document at once) and uses a novel sparse attention technique that yields 15× faster outputs even at these huge scales (cast.ai [3]). Kimchi's orchestrator intelligently delegates work: smaller, cheaper models handle routine coding steps, and it brings in M3 only for complex problems requiring that massive context or extra reasoning power. In side-by-side testing, this dynamic multi-model approach produced equivalent or better code results compared to using a single premium model - while cutting overall AI computation costs by about 60% (cast.ai [4]). In short, careful orchestration of multiple AI models can significantly reduce expenses without sacrificing performance.

The takeaway for technology leaders is that an "AI model portfolio" strategy may become standard. Open-source models are rapidly closing the gap with proprietary ones in certain domains, offering what one Cast AI executive calls 'frontier performance and economics that work at enterprise scale.' By mixing model types and optimising usage, companies can avoid vendor lock-in and control costs, all while still tackling complex tasks with AI where it truly adds value.

References:

[1] cast.ai — https://cast.ai/press-release/minimax-m3-comes-to-kimchi/#:~:text=becomes%20a%20strategic%20question%20for,built%20Kimchi%20to%20give%20every
[2] cast.ai — https://cast.ai/press-release/minimax-m3-comes-to-kimchi/#:~:text=Cast%20AI%27s%20Kimchi%20Coding%20Becomes,announcement%20comes%20as%20model%20selection
[3] cast.ai — https://cast.ai/press-release/minimax-m3-comes-to-kimchi/#:~:text=benchmark%20based%20on%20real%20GitHub,to%20push%20the%20boundaries%20of
[4] cast.ai — https://cast.ai/press-release/minimax-m3-comes-to-kimchi/#:~:text=developer%20teams%20frontier,a%20serverless%20deployment%20on%20Cast

From prompts to reusable skills

One of the more unexpected developments in the past 48 hours wasn’t a new model or product, but a new way of thinking about how employees use AI. The days of letting every staff member or developer craft prompts from scratch may be numbered. Highlighted in marketing analytics firm Trust Insights’ industry newsletter this week was a 'Prompt-to-Skill' plugin for Anthropic’s Claude AI, which converts any well-crafted prompt into a reusable skill that other team members (or even other AI agents) can invoke as needed. In essence, it packages a successful prompt - along with context (like examples or scripts) - into a shareable, standardised function that can be plugged into different workflows.

Why does this matter? Many organisations have seen the chaos that arises from purely ad-hoc prompt engineering. Results can vary widely between users and attempts, and hard-won prompting tricks often live only in individual heads or scattered documents. By contrast, a 'skill' is more like a

mini-application: it’s documented, version-controlled, and can be invoked reliably by various users or even other programmes. In a recent test at Trust Insights, layering their proprietary 5P methodology (Prompt, People, Process, Platforms, Performance) onto a raw prompt turned a 44% success rate into 100% (academy.trustinsights.ai [1]). In plain terms, formalising and standardising your best AI prompts into well-defined skills can dramatically improve consistency and outcomes.

For senior leaders, this is a signal to capture institutional AI knowledge and make it reusable. Just as software libraries allow developers to reuse code, codifying effective AI prompts as shareable skills can multiply their impact across the organisation. It’s an example of moving from artisanal, one-off AI interactions towards a more scalable, enterprise-grade practice.

References:
[1] academy.trustinsights.ai — https://academy.trustinsights.ai/products/digital_downloads/prompt-to-skill-plugin#:~:text=same%20prompt,skill%20runs%20it%20through%20the

Shining a light on AI bottlenecks

Even as AI models get faster and more powerful, companies are finding that technology isn’t the only barrier to speed. In fact, according to one report, 95% of AI pilot projects still fail to reach production - largely due to 'the difficulty of integrating agentic workers into their processes' (www.celonis.com [1]). In other words, human and organisational bottlenecks - slow handoffs, unclear accountability, and a lack of feedback loops - can stall even the most promising AI initiatives.

This realisation is sparking new interest in AI-focused process intelligence. In an analysis this week, Intellyx highlighted Bloomfilter, a platform that provides process-level observability across a company’s software delivery pipeline to help optimise AI adoption from a time-to-value and cost perspective (intellyx.com [2]). By capturing both high-level productivity metrics and developer-level activity data, such tools can uncover where AI projects are getting stuck or incurring waste, thereby revealing why returns might be lagging. For example, some teams that rushed to implement AI in every workflow found their token usage costs skyrocketing without commensurate gains (intellyx.com [3]). A platform like Bloomfilter can flag these patterns early, enabling leaders to intervene and adjust course before small inefficiencies turn into big setbacks.

This push for visibility extends into live operations as well. As noted earlier, AWS’s new OpenSearch MCP Apps allow an AI agent to directly access logs, metrics, and traces during incident response - another form of making the 'black box' of AI more transparent to human teams. Whether it’s during development or after deployment, the message is that you can’t manage what you can’t see. To truly reap the benefits of autonomous workflows, leaders should treat AI initiatives like any other critical business process: instrumented, monitored, and continually optimised.

References:
[1] www.celonis.com — https://www.celonis.com/news/press/bloomfilter-unveils-agent-miner-app-to-observe-govern-agents#:~:text=Development%20Enterprises%20are%20rushing%20to,app%20solves%20this%20coordination%20breakdown
[2] intellyx.com — https://intellyx.com/2026/06/14/bloomfilter-process-and-context-optimization-for-ai-and-agentic-development-adoption/#:~:text=are%20making%20things%20better%2Ffaster%2Fcheaper%20with,We%E2%80%99ve
[3] intellyx.com — https://intellyx.com/2026/06/14/bloomfilter-process-and-context-optimization-for-ai-and-agentic-development-adoption/#:~:text=can%20drag%20down%20long,as%20making%20complex%20decisions%20or

Key Statistics

- 180,000+ - Businesses that use 1Password’s vault to protect credentials ([www.morningstar.com](https://www.morningstar.com/news/business-wire/20260615920466/1password-introduces-credential-broker-building-a-secure-credentialing-layer-for-humans-machines-and-ai-agents#:~:text=Building%20on%20the%201Password%20vault,machine%20workloads%2C%20and%20AI%20agents))

- 2.5× - Reduction in coding costs using Cast AI’s multi-model agent vs. a single large model (with no drop in quality) ([cast.ai](https://cast.ai/press-release/minimax-m3-comes-to-kimchi/#:~:text=developer%20teams%20frontier,a%20serverless%20deployment%20on%20Cast))
- 4.5× - Faster software deployment in Amazon’s AI-driven development (a project done in 76 days vs. an original 12 - 18 month estimate) ([www.develeap.com](https://www.develeap.com/news/aws-weekly-roundup-aws-finops-agent-in-preview-gemma-4-on-be/#:~:text=Jun%2015%2C%202026%20AWS%20Weekly,AI%20infrastructure))
- 95% - Share of AI pilot projects that never reach production due to integration and process issues (MIT study) ([www.celonis.com](https://www.celonis.com/news/press/bloomfilter-unveils-agent-miner-app-to-observe-govern-agents#:~:text=Development%20Enterprises%20are%20rushing%20to,app%20solves%20this%20coordination%20breakdown))
- 56 points - Improvement in task success rate (from 44% to 100%) by converting a one-off prompt into a structured AI skill with a 5-step framework ([academy.trustinsights.ai](https://academy.trustinsights.ai/products/digital_downloads/prompt-to-skill-plugin#:~:text=same%20prompt,skill%20runs%20it%20through%20the))

KEY TAKEAWAY

These developments show that deploying AI agents successfully requires more than smart models. Leaders must invest in secure credential flows, cost controls, standardised AI 'skills' and robust process oversight to turn agent automation into real business value.

Sources

1Password Credential Broker reduces secret sprawl through identity-based credential delivery - Help Net Security
https://www.helpnetsecurity.com/2026/06/15/1password-credential-broker-reduces-secret-sprawl-through-identity-based-credential-delivery/

AWS FinOps Agent is now available in preview - AWS
https://aws.amazon.com/about-aws/whats-new/2026/06/aws-finops-agent-preview/

Amazon OpenSearch Service launches MCP Apps for agentic observability - AWS
https://aws.amazon.com/about-aws/whats-new/2026/06/opensearch-agentic-observability-mcp-app/

Cast AI’s Kimchi Coding Becomes the First Autonomous Coding Agent to Offer MiniMax M3, Delivering Frontier Open-Weight Performance at a Fraction of the Cost
https://cast.ai/press-release/minimax-m3-comes-to-kimchi/

Bloomfilter: Process and context optimization for AI and agentic development adoption – Intellyx
https://intellyx.com/2026/06/14/bloomfilter-process-and-context-optimization-for-ai-and-agentic-development-adoption/

New Digital Download – Prompt to Skill plugin (Trust Insights 5P Framework)
https://academy.trustinsights.ai/products/digital_downloads/prompt-to-skill-plugin

Bloomfilter Unveils Agent Miner App to Observe & Govern Agents - Celonis Press Release
https://www.celonis.com/news/press/bloomfilter-unveils-agent-miner-app-to-observe-govern-agents