

Beyond the First Agents:

# Running AI as a system

How to scale an AI estate without letting cost, risk and operational complexity scale with it.

**Gareth Martin**

Chief Data & AI Officer



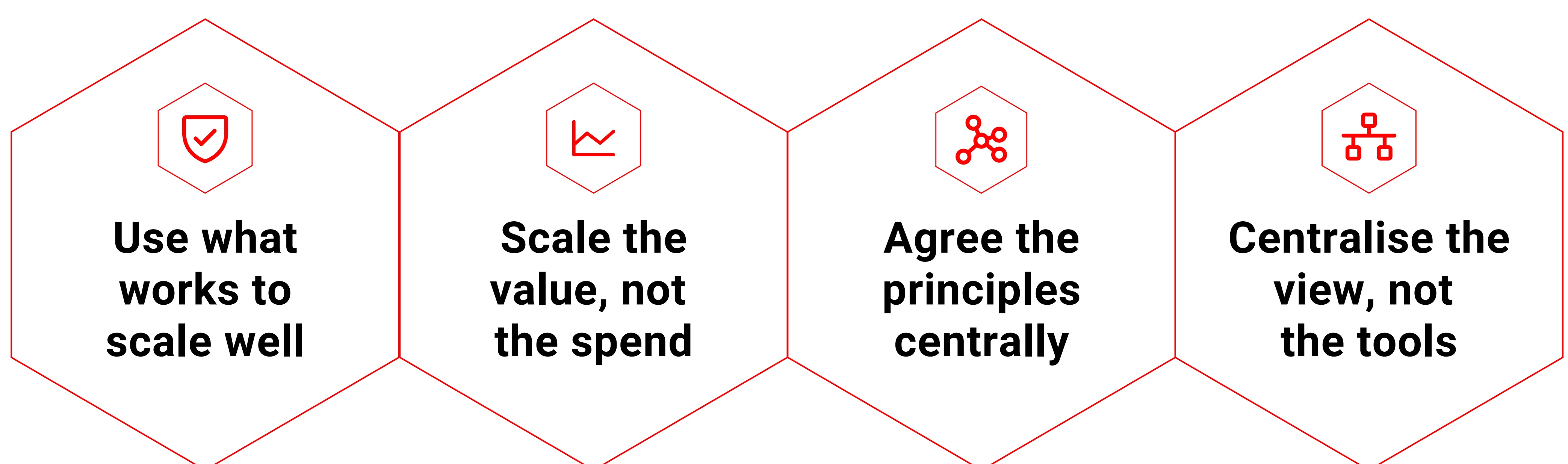
## Your first wave of AI agents shows what is possible. What you learn from them determines how effectively you can scale from there.

Once AI spreads across products, processes and teams, the challenge changes. One team builds against enterprise data while another switches AI on inside Salesforce. Before long, the question is not whether AI works, but whether you can see what exists, what it costs, what risk it creates and who owns it.

### That is how progress becomes sprawl.

The answer is not to force every agent onto one platform. It is control without unnecessary centralisation: define the principles and monitor the system centrally, and let the team that owns each tool manage the AI inside it.

We're focused here on custom, high-stakes agents operating inside core business processes. Four steps make that model work:





## Use what works to scale well

Most organisations will already have an initial group of agents running across different use cases. At that point, the question should not simply be what can we build next? It should be: **what have we learned from what is already running, and how can we use those lessons to scale more economically, securely and safely?**

That requires evidence across the first wave: which agents are delivering measurable value, which are being used, whether their costs are proportionate to the return and whether they are governed well enough to trust. From there, you can decide what to scale, what to improve and what to stop, as well as which patterns and guardrails are worth carrying forward.

Successful agents should leave behind reusable architecture, governance controls, monitoring patterns and a clearer understanding of how different categories of agent behave. Those lessons are what make the next phase of scale more repeatable and controlled. That is the first principle of scaling: reuse proven patterns and retune them for new use cases, rather than treating every build as bespoke.

### Five questions for scaling well

As the estate expands, keep five questions in view:

1

**Value:** Are the agents moving the business KPIs they were built to improve?

2

**Adoption:** Are people actually using them?

3

**Cost:** Are the costs proportionate to the returns, and is the model efficient enough to scale economically?

4

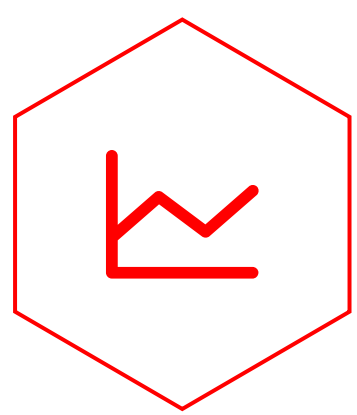
**Confidence:** Are they governed well enough to trust?

5

**Reuse:** What patterns, guardrails and monitoring from the first wave can be reused as the estate expands?

Demand is rarely the constraint. Marketing, HR, finance, supply chain, logistics and procurement will all have use cases they want to introduce or expand. The judgement is how to scale well, because moving too slowly can leave value on the table, while scaling without the right controls can multiply cost, risk and operational complexity.

With more at stake as the estate grows, a scalable AI capability should compound: every successful build should leave the organisation better equipped to approve, govern and operate AI at greater scale.



## Scale the value, not the spend

If each agent in the first wave costs £100,000 to build, that cannot remain the cost model as the estate expands. **Scale depends on reducing the cost of additional builds while keeping the run cost under control.**

In practice, that means reusing the same performance, governance and security patterns, **treating monitoring as configuration rather than a fresh engineering task**, and retuning existing agent configurations where the use case falls into a category you already know.

Running multiple agents introduces a second cost curve: the manage-and-operate cost rises as you add technologies to the estate, and rises again as usage volume increases. **More platforms mean more places to monitor consumption, and because AI usage can accelerate quickly, traditional month-end visibility is not enough.**

You'll already have budgets, approval thresholds and spending controls, and AI should sit inside that existing financial control model. The difference is that consumption can rise materially after the original approval, so the monitoring and intervention need to keep pace.

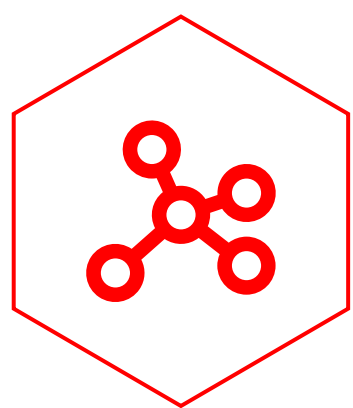
**This is also where you need to move from token-maxing to value-maxing.** The aim is not to consume more tokens or use the largest model or create the most technically elaborate response, but to achieve the required outcome at an efficient cost.

### Build the cost analysis around three questions



Those checks need to stay live after approval. An agent may start within budget, then become more expensive as more people use it, more workflows depend on it or a model change increases consumption. **Control needs to become more operational:** live monitoring against a threshold, with a pre-agreed response if consumption runs beyond it. Depending on the workload, that may mean alerting the owner, throttling usage or switching the service off.

Handled this way, AI spend becomes a managed P&L line with clear ownership and controls, rather than a growing cost that only attracts attention once the bill arrives.



## Agree the principles centrally, then let teams operate

In a large enterprise, AI ownership rarely sits neatly with one function. Technology, operations, strategy, risk and data protection all have a stake, which is why you need a cross-functional mechanism to agree the rules. **What you do not need is that group approving every individual agent, so monitor centrally, manage locally.**

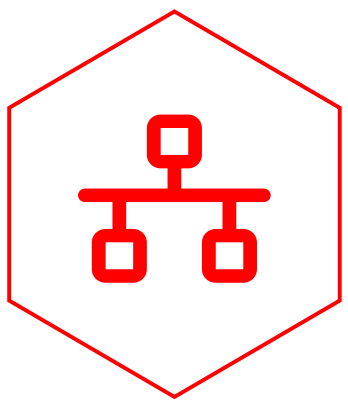
That starts by categorising agents according to what they do and the risk they introduce. An agent with access to PII or sensitive data that can then take action (e.g. contacting a customer without a human in the loop) will sit in a different risk category from a research agent. Automated pricing decisions may warrant tighter control, while content-generating agents should be assessed according to the risk in what they produce.

**Those categories also feed directly back into reuse, because agents in the same category can inherit common configuration and operating patterns.**

You will also need to agree on a set of principles that apply across the organisation. These principles are the common operating rules for AI: **they define what data can be used, which tools are permitted, how AI is registered and categorised, where agents should sit and who is responsible for managing them.** Every team should be able to use them to make day-to-day decisions within these agreed boundaries.

### Your Principles

- 1 Data**  
Only use data that is approved for AI. Sensitive, restricted or prohibited data must stay outside AI systems unless it has been explicitly cleared for that use.
- 2 Tools**  
Use approved AI tools and platforms only. If a tool has not been approved, it should not be introduced into a business process without review.
- 3 Embedded AI**  
Treat AI built into existing enterprise products as part of the AI estate. It can be used where the product already meets the organisation's required data protection and security standards.
- 4 Registration**  
Register every AI capability in use, including features embedded inside products the organisation already owns. If AI is generating content, making recommendations, supporting decisions or taking action, it needs to be visible.
- 5 Risk**  
Apply controls according to what the agent does, the data it uses and the actions it can take.
- 6 Placement**  
Build AI where it makes most sense for the process and the data. Keep system-specific agents close to the system that owns the workflow, and place cross-system agents close to the governed enterprise data they rely on.
- 7 Ownership**  
The team that owns the tool manages the AI inside it and remains accountable for how it operates.



## Centralise the view, not the tools

**Context does not travel automatically.** An individual agent may understand the data and definitions inside its own process, but as agents begin to work across systems, inconsistent context becomes an operating risk. If two agents are working from different definitions of the same metric, the technology can be functioning perfectly while the decisions are still wrong. **Cross-system agents therefore need to sit close to governed enterprise data, where the data, access and definitions they rely on are controlled consistently.**

For many organisations, Databricks often fits that cross-system layer because it is already the enterprise data platform, has a large talent pool around it, and provides governance over data access, lineage and auditability while supporting value-maxing over token-maxing.

One practical example is the pattern behind an AI Observability Command Center: collect logs and usage data from AI tools across the system, normalise them into a common data model, then use that governed layer to monitor adoption, spend, ownership, compliance and security risk. **This gives you one consistent source of operational data across a decentralised AI estate.** In practice, that means three steps: **collect the raw usage and audit data; refine it into a consistent model; then put it to work through dashboards, alerts, models and agents.**

### Build two layers of observability

- 01. Estate level:** Monitor what exists, who owns it, adoption, spend and risk across every AI tool.
- 02. Process level:** For each step within a workflow, identify the stages where performance can fail or spend can accumulate, define the measure at each stage, set the threshold that signals a problem, and decide what action follows (alert a person, pause the step or stop the process).

For example: imagine an agentic customer-acquisition process that designs campaigns, runs them, spends media budget and acquires customers. If acquisition starts falling, the final KPI tells you the process is underperforming but not why. The LLM may still be working perfectly; another stage may have failed, or one part of the chain may suddenly start producing German instead of English.

**Without enough monitors inside the process, you see the commercial result deteriorate without being able to pinpoint the cause.**

This is also a cost-control problem. In an agentic process, spend does not sit only in the model call; the workflow may be triggering offers, media spend or other downstream actions. **Step-level monitoring helps you identify where value and money are being lost before the final business metric exposes the problem.** The objective is to make intervention possible while the problem is still manageable.



## From agents to an operating system

Think of your AI system like a digital twin of a wind farm. Individual turbines operate locally, but each is visible as an asset in the system, reporting condition and performance so operators can see what needs attention across the whole estate.

The same logic applies to AI: **treat agents as assets, surfaced by risk, performance and the need for intervention.**

Agents remain close to the systems, teams and data they depend on, while the organisation provides the reusable patterns, shared principles, financial controls and central observability needed to keep performance, risk and intervention visible across the system.

**This is control without unnecessary centralisation.**

Once that system is working, the next question is where to point it.

The answer here is clear: **follow the money.** The highest-revenue areas, the highest-cost areas, underperforming processes with genuine upside and the journeys that shape customer experience are where the value is. Start with a controlled slice of a critical process, keep the right human intervention in place, then expand as confidence grows. And keep the system live. **An agent is not a project you finish and forget; it needs to be tuned and optimised as the market, model and process change.** The risks are not fundamentally new, but agents can act faster, so the consequences arrive faster too.

**Scaling AI is not about simply multiplying the number of agents in use.** It is about creating a system in which AI becomes progressively cheaper to build, easier to govern, simpler to monitor and safer to operate as the estate grows, without forcing the innovation itself into one place.

# We're Manuka. We build AI on Databricks. Let's build together.

📍 UK – Manchester – London – Liverpool

🌐 [linkedin.com/company/manukaai-emea](https://www.linkedin.com/company/manukaai-emea)

📍 ANZ – Sydney – Auckland

🌐 [linkedin.com/company/manukaai-anz](https://www.linkedin.com/company/manukaai-anz)

📍 US – California – North Carolina – Texas

🌐 [linkedin.com/company/manukaai-us](https://www.linkedin.com/company/manukaai-us)

🌐 [www.manuka-ai.com](https://www.manuka-ai.com)

