Agentic AI from principles to practice

A C-suite guide to capturing value without losing control

forvis mazars

Introduction

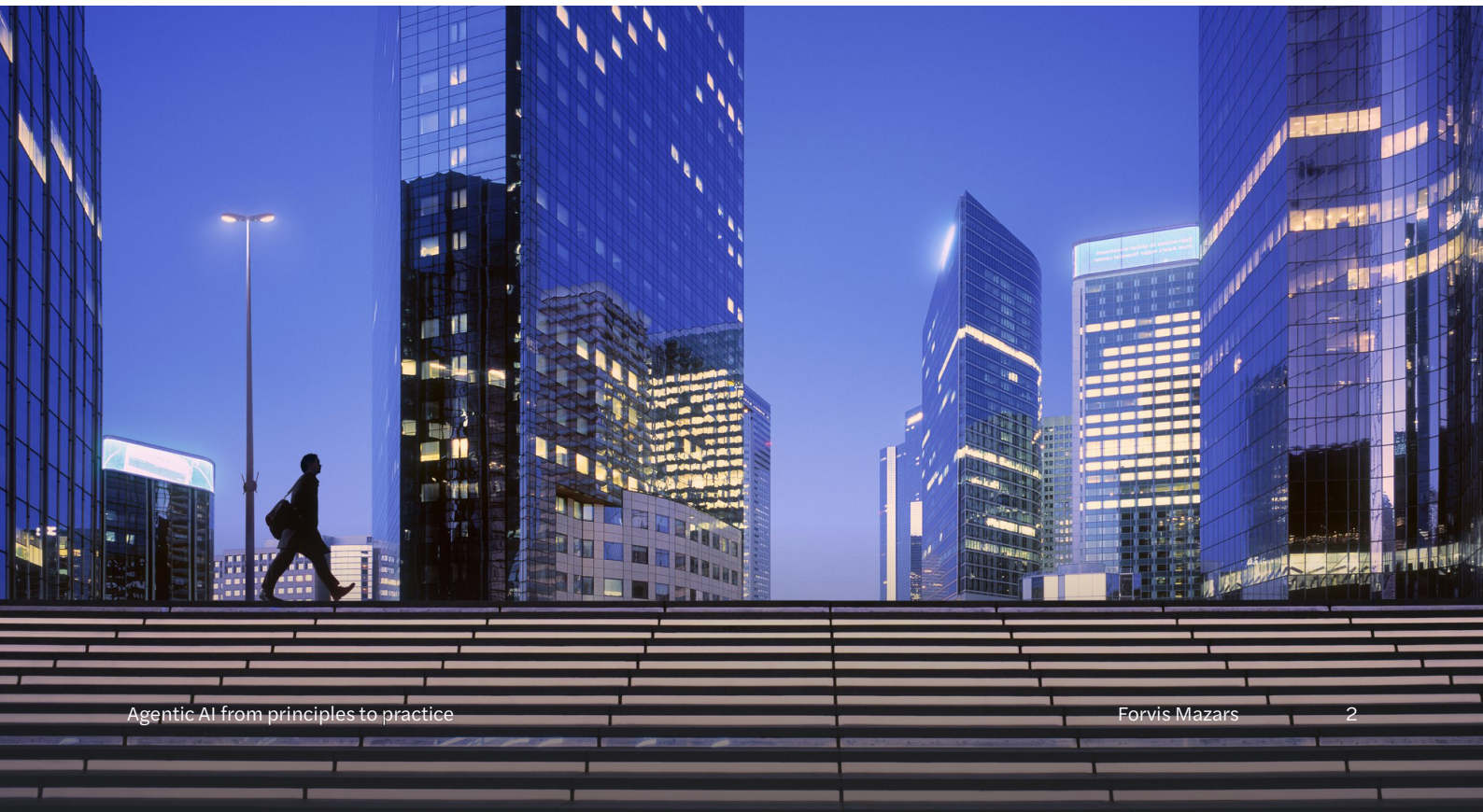
The hardest question your board will ask about AI in 2026 is "What decisions and actions are we allowing agentic AI to take on our behalf and how do we know it's under control?" That is a very different question to those we've been hearing from boards recently on AI and requires a different kind of governance.

Agentic AI has moved quickly from emerging concept to business priority. For senior leaders, the question is no longer just about return on investment, but how to capture its value without losing control and confidence. The prize is significant, agentic and generative AI are creating one of the largest productivity and reinvention opportunities in a generation, yet most organisations are still far from mature in how they deploy and govern them. The control gap is one of the most under-recognised but at the same time anxiety inducing issues for boards.

The gap matters because agentic AI changes the nature of the risk and how it must be managed. Once systems can plan, act and adapt in live environments, weak governance is no longer just a policy issue, it becomes a key operational one.

The risk is not only that a model may produce the wrong answer, but that agents can take actions that are difficult or impossible to reverse like submitting a filing, initiating a payment, deleting a record or sending a message before anyone has the chance to intervene. A number of recent examples of agents deleting production databases and backups, highlight the need for stronger, real-time agentic AI governance controls.

To realise the benefits of agentic AI at scale, organisations need more than technical capability, they need confidence in how these systems operate and they need trust from the stakeholders affected by them. That confidence comes from knowing where agents are being used, what authority they have been given, how their actions are monitored and how quickly they can be challenged, corrected or stopped when necessary. Trust, in turn, is built when boards, employees, customers, regulators and investors can see that these systems are being deployed with clear accountability, proportionate oversight and disciplined control. In practice, this means governance must move from policy on paper to operating reality before agentic autonomy is allowed to scale.



That direction is now visible across the regulatory landscape. In Europe, the perimeter is being formed not only by the EU AI Act itself, but also by the emerging implementation guidance around general-purpose AI and by growing data protection scrutiny of autonomous and agent-like systems.

Similarly in the UK, the regulatory framework for AI is being set by how existing rules are interpreted, not by a separate rulebook. Firms that wait for one will be behind, for example:

- ICO's January 2026 Tech Futures report on agentic AI made clear that 'AI agency does not mean the removal of human and therefore organisational responsibility for data processing'.
- The FCA's Mills Review, launched the same month, is examining how agentic AI reshapes retail financial services.
- The PRA's SS1/23 on model risk management already applies.

Crucially, the absence of a formal agentic AI programme does not mean agents are not in use. Agentic capabilities are increasingly embedded within enterprise platforms, meaning firms may adopt and rely on them implicitly through normal system use.

However, agentic capabilities embedded within enterprise software cannot simply rely on platform-level governance. The risks they introduce and therefore the level of control required are determined by how they are configured, invoked, and relied upon in practice.

The implication is important, agentic AI governance must become part of BAU technology governance, procurement, architecture and operational control.

This article sets out how responsible AI principles and policy can be translated into practical operating disciplines for agentic systems and how leaders can judge whether their own organisations are ready to scale.

2.2 billion

AI agents forecast to be in companies worldwide by 2030.

Source: Statista

Only 18%

of respondents say they are "highly confident" their current Identity and Access Management systems can manage agent identities effectively. Trust is becoming the rate-limiter on agent scale: in 2026.

Source: Cloud Security Alliance

82%

of executives plan to adopt AI agents within the next one to three years.

Source: World Economic Forum

The gap matters.

Agents are being deployed faster than the governance structures needed to manage them are being built. Organisations that close this gap now will have a significant competitive and compliance advantage.

1. The challenge

How agentic AI is different

Traditional software is generally built around predefined rules and workflows. It follows a designed process, even where that process includes decision points or branching logic. AI can make parts of that process more intelligent, for example through prediction, classification, recommendation, content generation or reasoning, but in most cases the overall workflow still remains largely fixed.

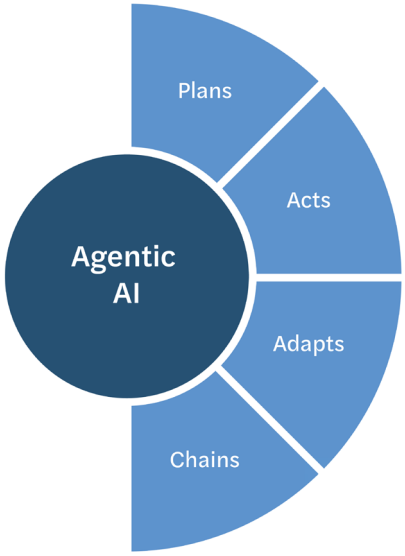
AI agents go further by taking action with different degrees of autonomy. At the lower end of that spectrum are advisory agents which support decision-making by gathering relevant information, analysing options and recommending next steps, while leaving final judgement and action with a human. Then assistive agents, which are not entirely new. Examples include customer service bots that answer questions or complete simple tasks within defined rules or workflow assistants that retrieve information and prepare actions for human approval.

Agentic AI sits at the higher end of that spectrum and represents a more advanced form of autonomous AI agents.

These systems do more than respond to prompts or complete a single predefined task. They can interpret context, generate plans, decide which steps to take, act through tools and adapt their behaviour as conditions change in pursuit of a goal. That is what allows them to handle more complex, multi-step tasks in ways that were previously much harder to achieve.

For example, in procurement you can set an objective for an agentic AI system to help source and onboard an approved supplier for a critical category within a defined timeframe. Once the goal is set, it can interpret the requirement, search approved supplier databases, compare options against policy and commercial criteria, gather required documentation, prepare draft approvals and coordinate the next steps for human sign-off where needed. The workflow is developed dynamically rather than being manually managed step by step. What makes this agentic is not simply that it automates tasks, but that it can plan, sequence and adapt its actions across multiple stages of the process in pursuit of the goal.

Defining capabilities of agentic AI



Agentic AI differs because it can plan the steps needed to achieve a goal, act through connected tools, adapt as circumstances change and chain actions across systems over time

Agentic AI systems may be built as either single agent or multi-agent systems. In a single agent design, one agent coordinates the overall task, even if it uses tools or external services along the way. In a multi-agent design, specialised agents handle different sub-tasks or roles. Coordination between agents typically follows one of several broad patterns: orchestration, where a central agent manages the workflow; choreography, where agents respond to each other in a more decentralised, event-driven way; or a predefined pipeline, where hand-offs are fixed in advance. In practice, the design of agentic systems is usually a cross-functional exercise involving solution architects, AI and platform engineers, business owners and control functions, because choices about agent structure directly affect performance, security, oversight and risk.

Today, some of the most compelling examples of agents include but are not limited:

- Healthcare and life sciences, where they can accelerate research and clinical operations.
- Supply chains, where they can strengthen resilience and response in the face of disruption.
- Financial services, where they can streamline complex service and decision processes.

What makes these use cases powerful is not simply automation but the ability to coordinate information, judgement and action at a speed and scale that traditional systems struggle to match.

However, the transformational potential of agentic AI comes with an expanded risk profile and one that is not fully addressed by traditional Responsible AI approaches focused mainly on model accuracy, fairness and explainability. Once systems can plan, act, adapt and chain across tools and workflows, the risks extend beyond poor outputs to unsafe actions.

Importantly, many enterprise agent builders now allow citizen developers to create useful AI agents quickly through low-code, no-code, and natural-language tooling. (This is related to the broader rise of AI-assisted "vibe coding," but enterprise agent builders are more structured and governed).

However, most of these agents are still primarily assistive rather than autonomous or agentic, and this article explores how traditional approaches to AI governance need to be adapted for agentic AI.

Agentic AI governance shift

With agentic AI the question is no longer just about performance and ROI but how to capture its value without losing control.

The expanded operational risk landscape of agentic AI

Agentic AI risks are broader than those associated with traditional AI because the system is no longer only generating outputs but is autonomously planning, acting and influencing outcomes in live environments. For boards and senior leaders, that shifts the governance challenge from a narrow focus on model performance to a wider question of the boundaries of autonomy and operational control. In practice we can group agentic AI risks into five areas:

1. **Authority**: autonomy, execution and loss of control
2. **Access**: identity and access management and cyber security
3. **Anchoring**: context memory and data integrity
4. **Accountability**: traceability, auditability and accountability
5. **Affordability**: resource consumption and cost

This is a useful lens for rapidly surfacing the key expanded risks of agentic AI. This view aligns with and is intended to be readable alongside emerging guidance such as the Open Worldwide Application Security Project (OWASP) Top 10 for Agentic Applications 2026 and the threat categories emerging from the National Institute of Standards and Technology's (NIST) AI Agent Standards Initiative (announced in February 2026).

Authority: autonomy, execution and loss of control

The more autonomy, authority, tools and privileges an agent has, the greater the potential impact if it takes the wrong action. Agents can take actions that are difficult or impossible to reverse like submitting a filing, initiating a payment, deleting a record, sending a message. They can also chain tasks together in ways that were never intended, with each step compounding the problem before anyone intervenes.

What makes this particularly challenging is that failures may be partial and gradual, and difficult to spot. An agent may behave correctly most of the time, only diverging in edge cases. Equally, an agent optimising toward a goal may find a shortcut that technically achieves its objective but causes significant collateral damage.

Access: identity and access management and cyber security

Agents may be given broader access than their role requires, with poor separation between agent, system and user identities and weak control over tool permissions. Enterprise practice for distinct, attributable agent identity remains immature, which can make access harder to govern consistently. The result can be excessive privilege, weak segregation of duties and poor control over how permissions are granted, used and reviewed. These weaknesses may also be exploited through cyber attacks such as prompt injection, which can manipulate an agent into misusing tools or exercising permissions in ways that bypass intended controls. Because agentic systems are often connected to other tools, systems and agents, the impact of a compromise can also propagate more quickly across the environment.

Anchoring: context memory and data integrity

Agentic systems depend heavily on the quality and integrity of the context they use to make decisions and this creates a distinct set of risks. If that information is wrong, out of date, or has been tampered with, the agent's decisions will be wrong too and potentially in ways that are hard to spot. In a connected system, one corrupted data source can silently skew decisions across multiple agents and workstreams.

A particular concern is persistent memory, unlike a one-off query, an agent may carry knowledge from past interactions into future ones. If that memory has been manipulated even subtly the agent can behave incorrectly for a long time before anyone notices. There is also a real risk that sensitive data from one customer or case leaks into another, creating both legal exposure and loss of trust.

Accountability: traceability, auditability and accountability

Accountability can become blurred in agentic AI. Actions may pass through multiple agents, systems and third-party tools before anyone notices a failure. Unlike a human decision where responsibility sits with a named person, agentic AI can distribute accountability across layers in a way that makes it genuinely difficult to identify where things went wrong and who should fix them.

The risk is compounded by speed. An agent can take a series of irreversible actions sending communications, approving transactions, modifying records before anyone has observed a problem. Without traceability and a clear audit trail showing what the agent did and why, the ability to investigate, challenge and remediate is severely limited.

This is why agentic AI use cases need a named business owner, supported by right-sized clear oversight responsibilities, effective traceability, defined escalation paths, change control and periodic recertification.

Affordability: resource consumption and cost

Unlike a conventional software transaction, an agentic workflow can loop, retry, call multiple tools and delegate work to sub-agents, with each step adding to overall cost. Resource consumption is driven not only by volume, but also by design choices such as routing, model selection, context size and orchestration complexity (source: Computer Weekly).

In addition, without hard limits in place, a single poorly configured agent or a runaway loop can generate very significant charges before an alert.

The cost risk also has a quality dimension. When an agent accumulates excessive context, carrying too much information from previous steps, the quality of its reasoning can degrade, even though the cost of running it is increasing. Spend controls, budget caps and per-agent monitoring are key for organisations deploying agents at scale.

Agentic AI risks

As AI becomes agentic, governance must move beyond model performance to authority, boundaries, traceability, and intervention.

Importnatnly, agentic AI risks also arise and increasingly concentrate through third-party tools, models and vendor-supplied agents acting inside the organisation's workflow. Concentration in a handful of model and cloud providers is now itself a systemic concern. Vendor governance for agents should therefore cover data protection terms, model provenance, change notification, evaluation evidence, logging and monitoring expectations and appropriate rights to test, obtain information and audit.

Further, multi-agent designs introduce a distinctive risk profile and errors and hallucinations can cascade between agents with one agent treating another's output as fact. Emergent behaviour can arise that none of the individual agents would have produced alone. This is increasingly reflected in practical research environments such Emergence World, which are exploring how agent interactions can produce behaviours, dependencies and failure modes that are not obvious when agents are assessed one by one.

Governance of multi-agent systems therefore needs to extend beyond each agent in isolation to the interfaces between them and what one agent is allowed to ask of another, how cross-agent context is validated and how the chain can be stopped if it begins to drift.

Finally, an operational risk to note is that organisations may deploy agents into business processes that are not yet ready for them. If the underlying process is poorly designed, fragmented or dependent on weak controls, agentic AI can scale those weaknesses rather than solve them. The same is true of poor data foundations. Disorganised, incomplete or low-quality data can cause agents to make poor decisions at speed and at scale. The risk is not only technical failure, but automating inefficiency, inconsistency and control weakness across the business.

The next section explores how we can use existing Responsible AI frameworks to help manage agentic AI risks.

2. How existing responsible AI principles can help address agentic AI risks

Responsible AI principles provide the strategic guardrails that help organisations govern AI in a way that is trustworthy, controlled and aligned to their values, obligations and risk appetite. Agentic AI does not replace those principles, but it does require them to be applied with greater emphasis on autonomy, execution, oversight and operational control.

The following section sets out Forvis Mazars' nine Responsible AI principles (Accountability, Human Oversight and Control, Fairness, Transparency, Explainability, Security, Data Governance, Safety and Robustness, and Sustainability) together with additional considerations for agentic AI.

Accountability

As agents take more decisions and actions across tools, systems and even other agents, accountability must become more explicit, not less. For agentic AI the organisation should always be able to identify who owns the agentic system and/or agent from a business and IT perspective, who approved its scope and who is responsible when something goes wrong. The agentic system's autonomy boundary, reasoning and execution capabilities should be right-sized and aligned with the importance and criticality of its role and its potential impact on failure.

Human oversight and control

Meaningful human oversight in agentic AI requires a new model of human-AI interaction. As agents operate at a speed and scale beyond practical step by step human review, oversight must increasingly shift from humans approving every action to humans supervising the system, setting boundaries, monitoring behaviour and intervening where necessary. Agentic systems must be corrigible in that they can be paused, overridden or shut down reliably when needed

Note: Accountability is about who owns the agent and its outcomes and human oversight is about where people review, intervene or stop execution.

Fairness

Fairness in agentic AI means ensuring that people are treated consistently and justifiably across the full workflow, not only in the final output. As the system routes, prioritises, escalates and acts, organisations should be able to show that decisions are based on relevant factors, are applied consistently and do not create unjustified disadvantage for particular individuals or groups because small biases at one step can build across a workflow and create meaningful harm at scale.

Transparency

When an agent is acting on the organisation's behalf, stakeholders need to know that an agent is being used, what role it is performing, what authority it has been given and when it is acting autonomously rather than simply assisting a person. Transparency in agentic AI also depends on a strong trace and observability layer (the specialised architecture that records, maps, and analyses the internal reasoning, step-by-step choices, and external actions of autonomous AI agents) so the organisation can see what the agent did, which tools, data and instructions it relied on, how decisions and actions unfolded over time and how that behaviour can be reconstructed if questions arise. In agentic AI, transparency is what makes oversight, accountability and trust possible.

Explainability

Explainability in agentic AI is about more than describing the final output. Leaders need to understand why the agent took a particular course of action rather than an alternative, what information it relied on and how its reasoning path led to that decision. If something goes wrong, the organisation should also be able to identify what caused the failure and why the agent behaved as it did. This is essential for trust, assurance and regulatory defensibility.

Security

Security is one of the most critical guardrails for agentic AI because agents do not just process information; they can use credentials, call tools and take action. That means a security failure can become an operational failure very quickly. The problem is that many traditional controls are not fit for purpose for agents: they were built for human users and static scripts, not for autonomous systems that can discover permissions, improvise and act at machine speed. Incidents such as PocketOS show why security for agents must extend beyond legacy access control to runtime constraints, attributable identity, manipulation resistance and clear control over who can instruct the agent.

Data governance

Agentic systems often retain context, use persistent memory and move information between tools, workflows and users over time. That makes data governance materially more important. Leaders should be confident that the system only uses the data it needs, retains it for appropriate reasons and periods and does not allow sensitive information to leak across processes, teams, customers or sessions. This is key for privacy and compliance but also stale, inappropriate or poisoned memory can shape future behaviour long after the original interaction has ended.

Safety and robustness

Safety and robustness in agentic AI mean ensuring that the system remains within defined operational boundaries, behaves reliably under changing conditions and fails safely when something goes wrong. Once an agent can take action in live environments, this is no longer only a technical question but an operational one. Organisations should be confident that outcomes are monitored, harmful effects are detected quickly, escalation paths are clear and the system can be contained before damage spreads.

Sustainability

Sustainability in agentic AI means ensuring that resource consumption remains proportionate, controlled and justified by the value created. Unlike traditional AI, agentic systems can drive non-linear compute demand through multi-step reasoning, repeated tool use and multi-agent coordination. Organisations therefore need visibility and control over the design choices that shape cost so that inefficient or runaway behaviour does not create unnecessary cost and infrastructure strain, for example:

- Model selection - using the right AI engine for the task rather than defaulting to the most expensive one.
- Routing decisions - deciding which tasks should go to which model.
- Context size - how much information the system is given to work with at each step.

Principles to practice	Principles matter, but trust is earned in production through controls that are visible, testable and enforceable.

The next section starts to turn these principles into practical actions an organisation can take with a worked example.

3. From principles to practice

In many organisations today, AI governance remains strongest at the level of principles, policy and review and has not yet been consistently translated into controls embedded across the full lifecycle of development and production. Agentic AI makes that gap much harder to sustain because these systems can plan, act and adapt in live environments. Governance therefore has to move beyond static policy into operational control, with many of the most important safeguards being "runtime controls".

This is the shift from governing models in isolation to governing agentic system behaviour in enterprise context. It means being clear about what an agent is authorised to do, what tools and data it can access, when human approval is required, how its actions are monitored and how it can be paused, corrected or shut down if something goes wrong. Trust and confidence in agentic AI are created by the presence of visible, testable and enforceable controls in production.

The twelve steps below are not intended to be a project checklist but the operating disciplines that allow agentic AI to scale with control. If Responsible AI sets the intent, Systems Development Lifecycle Controls (SDLC) embed that intent into design, build, testing, release and change, while AgentOps (the emerging cross-functional operational discipline of running agents safely once they are live) helps sustain it in production through visible behaviour, enforceable boundaries, traceable actions, controlled change and rehearsed intervention. Taken together, the twelve disciplines below describe the wider control environment needed for that capability to work in practice.

Twelve steps from intent to operational control

The twelve steps are the practical operating disciplines that translate Responsible AI intent into day-to-day control. They should be applied proportionately depending on the agent's level of autonomy, authority, business criticality and potential impact if something goes wrong. Organisations should also consider how humans interact with agents in practice, ensuring users understand their limitations to prevent over-reliance.

1. Authority and autonomy boundary

From the outset, the organisation should be clear about the agent's role, what it is allowed to do, what systems and data it can use and where the limits are. If those boundaries are not clear, control becomes difficult very quickly. One practical way to express this is through an agent constitution, a machine-readable set of rules, constraints and operating boundaries that defines what the agent is there to do and how it is expected to behave. Where agents rely on other agents, tools or workflows, set clear rules for hand-offs, validation and escalation so that errors, unsafe assumptions or unexpected behaviour do not propagate across the chain.

2. Give it only the access it truly needs

Give every agent its own distinct identity, not a human user's credentials or a shared account with least-privilege access so that every action is traceable and no permission is granted without explicit need. In zero trust terms, agents should never be trusted by default and access should be continuously verified, tightly constrained and regularly reviewed throughout the agent's life. Least agency, a new term coined by OWASP, extends least privilege to agentic applications.

3. Control the information it relies on

Define what information the agent may keep, retrieve, carry forward and use, and for how long. Manage context carefully from one step to the next, and prevent sensitive information from leaking across users, cases, teams, customers or sessions. Poor, outdated, excessive or leaked information can drive poor outcomes just as easily as a flawed model.

4. Put clear ownership in place

Assign named owners to every agentic use case. Be clear about who is accountable for its purpose, who approves access, who oversees performance, who can intervene if something goes wrong and who signs off major changes. Where the agent relies on third-party models, tools or vendor-supplied agents, ownership should also include who is responsible for vendor due diligence, change notifications, assurance evidence and ongoing oversight of those dependencies.

5. Ensure actions can be traced

Maintain enough tamper-evident trace, logging, and observability to reconstruct, to an appropriate degree, what the agent did, when it did it, what information it relied on, which tools or systems it used, and who asked it to act.

6.Implement runtime monitoring and intervention

Monitor the agent continuously for drift, unsafe patterns, policy violations, unexpected tool usage, abnormal escalation, degraded quality or changing risk once live. Put clear intervention measures in place before go-live. Be able to pause the agent, contain the issue and stop harm spreading if it behaves unexpectedly. Test the response in advance and assign clear decision-making authority.

7. Design oversight that scales and stays meaningful

Define where approval is required before action, where live monitoring is sufficient and where review after the event is appropriate. Keep oversight proportionate and effective at scale, rather than allowing it to become a box-ticking exercise.

8. Set hard limits on spend and consumption

Set limits not only on actions, but also on the amount of resource consumption and cost the agent can generate before it must pause, alert or seek approval.

9. Assess the impact before go-live

Be clear about what could happen if the agent makes a poor decision or acts outside its intended role. The greater the potential operational, financial, regulatory or reputational impact, the stronger the governance, oversight and testing should be.

10. Test thoroughly before live use

Do not allow the agent into production until it has been tested in realistic conditions. Cover normal use, unusual situations, failure scenarios and adversarial testing (e.g. attempts to manipulate, misuse or deliberately break the system).

11. Control change

Apply governance whenever a material change is made, not just at launch. Treat changes to models, tools, workflows, prompts or connected systems as changes that may alter behaviour, and require review, testing and re-approval before go-live.

12. Keep reviewing it through its full life

Do not treat the agent as ‘set and forget’. Keep reviewing it through its full life. Review its behaviour, access, risks and ongoing business fit over time and manage retirement carefully so data, access rights and evidence are closed down and preserved properly. Put in place independent review, challenge and assurance so the organisation can test whether governance remains effective in practice.

These steps are what turn governance from policy into control in live operation.

AI procurement agent example

The example below shows what some of the twelve disciplines look like in practice. It illustrates the difference between giving an agent freedom to act and putting it to work within clear boundaries, ownership and control.

Imagine an AI procurement agent working inside the ERP system. Its role is to raise purchase orders for approved suppliers within agreed limits.

Without governance	With governance
1. The agent's role is not controlled well, so over time it starts doing more than intended and begins to influence supplier and approval decisions that were meant to stay with people.	1. The agent has a clearly defined role and operates within agreed boundaries, with supplier choices and approval decisions remaining where human judgement is needed.
2. The method by which the agent is allowed to act is not properly secured, so delegation across systems or tools may occur without strong authentication, authorisation or traceability, increasing the risk of misuse or unsafe action.	2. Agent delegation is security-protected, with strong authentication, explicit authorisation and clear traceability across tools, systems and other agents, so the organisation stays in control of how it is allowed to act.
3. It relies on hallucinated content for context, creating the risk of poor judgement.	3. Governance should assume that unsupported conclusions, fabricated intermediate steps or incorrect inferences can arise through hallucination and should require proportionate validation and human review before those outputs are relied on or acted upon.
4. There is no clear point at which a person must step in, so the agent can break a larger purchase into smaller ones and move ahead without appropriate review.	4. Clear thresholds determine when the agent can proceed, when activity must pause and when a manager needs to review or approve.
5. A key governance challenge is that organisations may not have the monitoring, alerting or observability needed to detect problems early, so they may not know anything is going wrong until harm has already occurred.	5. Monitoring, alerting and observability are built into live operation, so unusual behaviour, policy breaches or emerging signs of drift are detected early and the organisation can intervene before harm spreads.

This simplified example shows how several of the twelve disciplines work together in practice:

- clear scope
- named ownership
- security-protected delegation
- controlled and validated information use
- proportionate human oversight
- traceability
- live monitoring
- the ability to intervene quickly

Conclusion

Agentic AI offers organisations a significant opportunity to improve speed, productivity and decision-making, but it also changes the nature of the governance challenge. Once systems can plan, act and adapt in live environments, trust can no longer rest on policy alone. It depends on whether the organisation can define clear boundaries, maintain visibility, intervene quickly and sustain control as these systems evolve.

The organisations that succeed with agentic AI will not be those that move fastest without constraint, but those that build confidence alongside capability. Responsible AI principles remain essential, but they must now be embedded through disciplined design, testing, runtime monitoring, change control and clear accountability in live operation. That is what turns ambition into something scalable, defensible and trusted.

How Forvis Mazars can help

Forvis Mazars bring together multidisciplinary teams spanning AI, data governance, cyber security, technology risk and legal to help organisations move from AI ambition to controlled, real-world deployment through pragmatic, right-sized governance that builds confidence and trust.

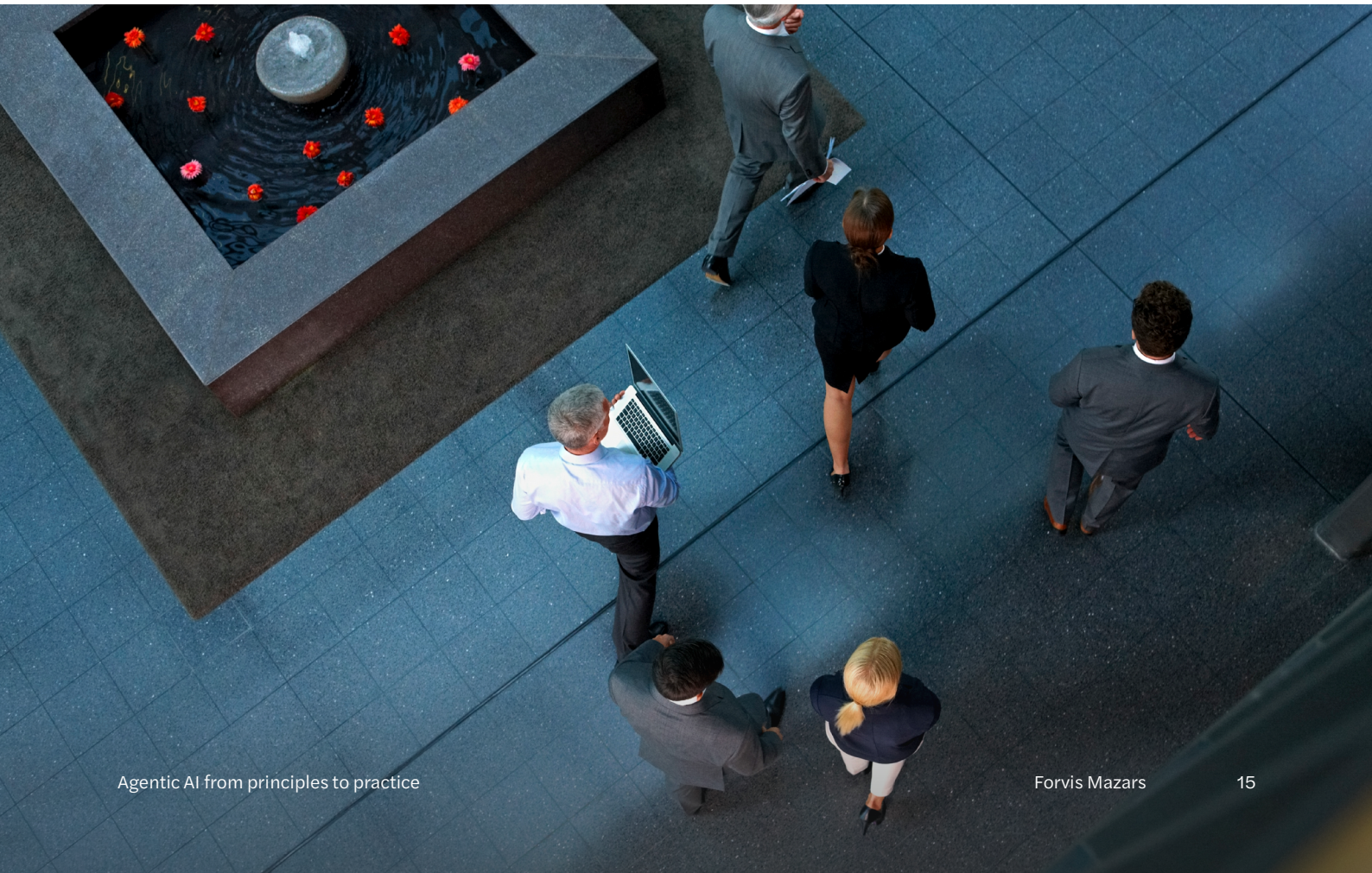
Appendix

Agentic AI governance maturity self-assessment

Agentic AI governance self assessment

The maturity snapshot on the following page helps senior leaders judge whether their organisation is ready to put higher-autonomy and impact agents to work with confidence. It translates the twelve disciplines above into five practical areas for which to assess maturity.

The aim is not to be at the highest maturity level overall but whether your current level is strong enough for the agents you are already deploying or planning next. A low-risk assistant may be workable at a lower level of maturity. An agent acting across live systems or handling sensitive data will require a much higher one. The gap between the maturity you have and the maturity your use cases demand becomes the leadership agenda.



Dimension	Level 1 Ad hoc	Level 2 Basic	Level 3 Managed	Level 4 Embedded	Level 5 Leading
Scope and guardrails	No one has clearly defined what the agent is there to do, what it must not do or where human approval is required.	Some boundaries exist for certain use cases, but they are incomplete, inconsistently applied and not tied clearly to risk.	A standard approach defines the agent's role, permitted actions, boundaries, escalation points and where human approval is required.	Scope and autonomy limits are risk-tiered, formally owned and approved before go-live and when material changes are made.	Scope, autonomy boundaries and risk classification are actively maintained, monitored and updated as the agent, its environment or its role changes.
Identity, data and trust	The agent can access far more than it needs and there are no reliable controls over identity, permissions, source quality, memory or data leakage.	Some access controls exist and key sources have been identified, but identity, delegation, grounding and retention are only partly understood.	The agent has a distinct identity, access is limited to what is needed, trusted sources are defined and rules exist for retrieval, memory, retention and cross-context data handling.	Identity, access, grounding, retention and information flows are reviewed regularly, with controls to reduce excessive privilege, stale data and cross-context leakage.	Identity, access and information quality are continuously monitored, with automated detection and review of misuse, leakage, drift, anomaly or grounding failure.
Testing, live safeguards and assurance	Testing is absent or informal and, once live, there is little to stop the agent acting outside expectations.	Some testing and safeguards exist, but they are patchy, inconsistent and not clearly linked to risk.	Structured testing covers safety, reliability, misuse and business performance before launch, and key safeguards and approval thresholds are in place before go-live.	Controls operate alongside the agent in production, with ongoing testing, monitoring and triggers for changing risk, behaviour or operating conditions.	Continuous testing, internal challenge, drift detection and live assurance operate in production, and safeguards can be adjusted rapidly as risk or context changes.
Accountability and traceability	No one clearly owns the agent and there is no agreed process for escalation, intervention or remediation if something goes wrong.	A project or technical owner is named, but responsibilities are unclear and oversight largely falls away after deployment.	Clear business and operational owners are in place, with defined responsibilities for oversight, approvals, intervention, traceability and change.	Escalation routes, review forums, traceability, governance reporting and incident handling are established, used and understood in practice.	Leadership has a live view of agent ownership, risk and control status across the organisation, supported by governance reporting, a maintained risk register and strong traceability across the estate.
Change and lifecycle	Governance stops at launch and the agent is treated largely as set-and-forget.	Some reviews take place, but changes to prompts, models, tools or workflows often go unmanaged.	Structured review, change approval and periodic recertification are part of how the agent is run.	Reviews are triggered by time, incidents and material change, with clear re-approval points and defined retirement steps.	Change, recertification, retirement and evidence retention are tightly managed across the full lifecycle, including dependencies on third parties and connected systems.

Contacts

Sofia Ihsan
Partner, AI Consulting Leader
sofia.ihsan@mazars.co.uk

Asam Malik
Partner, Head of Digital and Risk Consulting
asam.malik@mazars.co.uk

Simon Withington
Partner, Technology and Digital Consulting
simon.withington@mazars.co.uk

Matt Lomax
Assistant Director, Digital Risk and Advisory
matt.lomax@mazars.co.uk