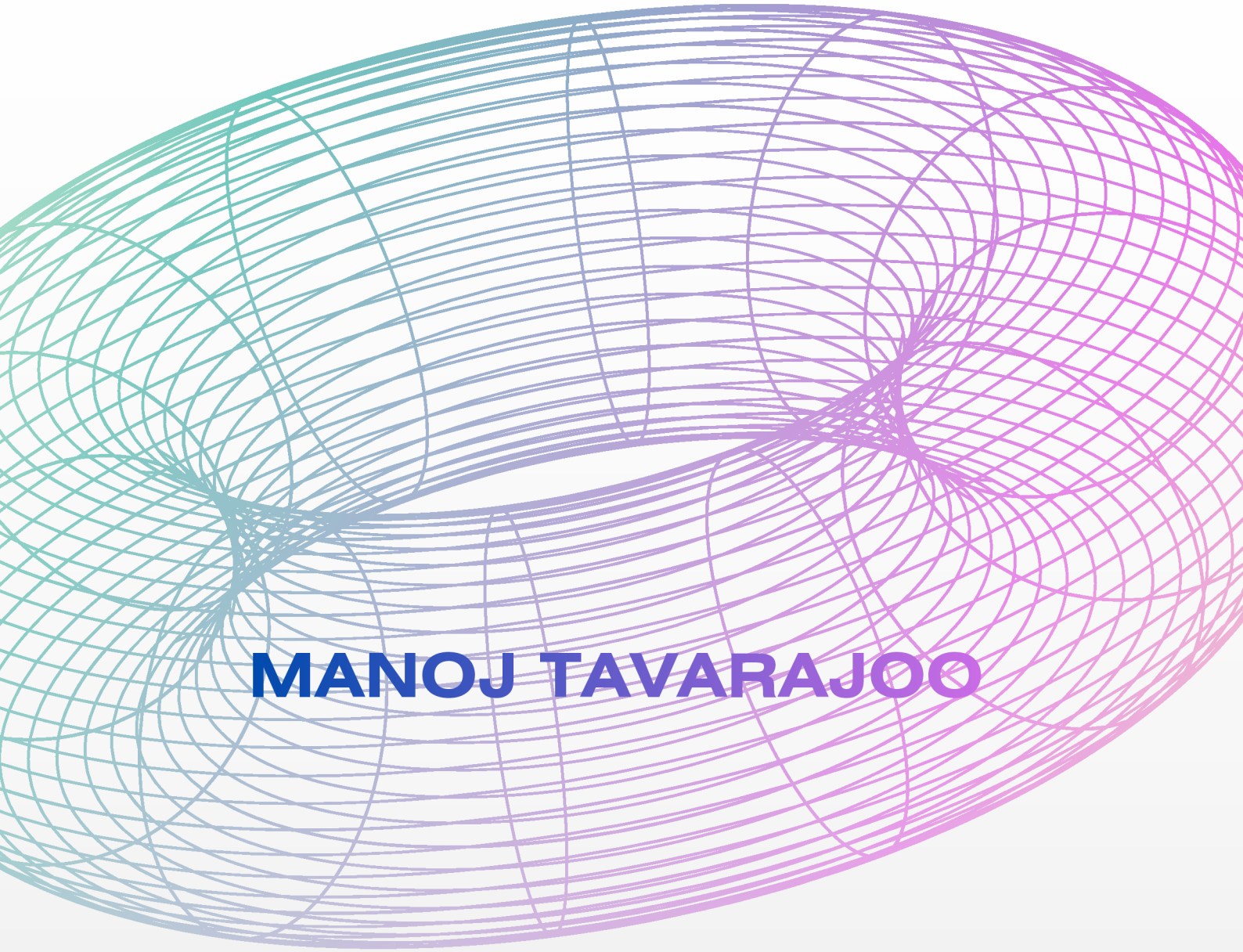
AI GOVERNANCE PAPER SERIES

MyConsultancy

GOVERNING AGENTIC AI

When AI systems move from advice to action



MANOJ TAVARAJOO

EXECUTIVE SUMMARY

The governance challenge changes fundamentally when AI systems can act, not merely advise.

Traditional AI systems generated predictions, classifications or recommendations that humans reviewed before making decisions. Generative AI copilots expanded this further by producing content, analysis and guidance that still depended on human approval and execution.
Agentic AI changes that model.

AI systems are now capable of initiating workflows, accessing tools, interacting with systems and executing actions on behalf of the organisation without requiring human approval for every step. The governance challenge is no longer limited to the quality of outputs. It extends to authority, permissions, escalation, intervention and accountability for autonomous action.

This changes the governance problem fundamentally.

Once organisations delegate operational authority to AI systems, governance must define what those systems are permitted to do, what systems they may access and when intervention becomes mandatory. It must also establish escalation thresholds, containment controls, evidence requirements and accountability for downstream actions.

This is already becoming an operational reality. Enterprise AI agents are increasingly being introduced through SaaS platforms, low-code automation tools and embedded vendor capabilities. Many organisations are therefore scaling agentic behaviour before establishing governance models designed to control it.

> The first large-scale enterprise agent deployments are arriving through SaaS enablement, not formal AI programmes.

This creates a structural governance gap.

Traditional AI governance focused on model approval, policy compliance and output review. Those approaches remain necessary, but they become insufficient once systems can initiate action across workflows, tools and systems.

This paper introduces a practical governance framework for agentic AI. It explains how organisations should govern delegated operational authority through autonomy boundaries, human oversight, escalation controls and containment models. It also examines the growing governance challenges associated with multi-agent delegation, SaaS-enabled agents and shadow-agent proliferation across the enterprise.

The paper argues that autonomy should not be treated as binary. It should be governed as a managed operating boundary.

The organisations that will succeed with agentic AI will not necessarily be those deploying the most autonomous systems. They will be those that define authority clearly, enforce operational boundaries and maintain evidence of control as autonomous execution scales.

> AI governance began as governance of models. It is becoming governance of delegated operational authority.

THE GOVERNANCE SHIFT

AI governance changes when systems move from generating outputs to taking delegated action.

Traditional AI systems predicted, classified or scored information. Humans reviewed the results and made decisions.

Generative AI systems expanded this further. Copilots could draft content, summarise information, generate analysis and support decision-making. Even then, human review remained central to execution.

> An AI system becomes agentic when it can initiate or execute action on behalf of the organisation without requiring human approval for every step.

Agentic AI changes the operational model.

The defining characteristic is not intelligence. It is delegated authority.

An agent may trigger workflows, access systems, retrieve or update records, coordinate across tools, communicate externally or orchestrate downstream agents without requiring human approval at every step.

This moves AI from advisory support into operational execution. That shift changes governance fundamentally.

The organisation is no longer only governing outputs. It is governing delegated operational authority.

THE FOUR-TYPE AI GOVERNANCE TAXONOMY

AI systems differ by what they do, the role humans play, and the governance challenge they create.

01 · TRADITIONAL AI
Predictive Systems

What it does
Predicts, classifies, scores or detects patterns in data.

Human role
Human makes the decision.

Governance focus
Model accuracy, fairness and explainability.

Primary Governance Risk
Incorrect or biased predictions lead to poor decisions.

Example Use Cases
Fraud detection, credit scoring, demand forecasting.

2. GENERATIVE COPILOT
Content & Recommendations

What it does
Generates content, analysis or recommendations for human use.

Human role
Human reviews and decides.

Governance focus
Output reliability, appropriateness and safety.

Primary Governance Risk
Misinformation or harmful / inappropriate output influences decisions.

Example Use Cases
Drafting emails, summarising documents, generating analysis or insights.

3. WORKFLOW AUTOMATION
Deterministic Execution

What it does
Executes predefined, deterministic rules and processes.

Human role
Human configures and monitors.

Governance focus
Process integrity, rule configuration and operational control.

Primary Governance Risk
Misconfiguration or rule failure disrupts operations.

Example Use Cases
Invoice processing, approval routing, IT service workflows.

4. AGENTIC AI
Autonomous Action

What it does
Plans, initiates and acts across tools, workflows and systems on behalf of the organisation.

Human role
Human delegates authority and intervenes by exception.

Governance focus
Delegated authority, autonomy boundaries and accountability for action.

Primary Governance Risk
Autonomous action outside approved boundary causes harm or loss.

Example Use Cases
Customer issue resolution, financial operations, supply chain orchestration.

GOVERNANCE INTENSITY & COMPLEXITY

Lower → Higher

KEY TAKEAWAY: As AI systems move from prediction to autonomous action, the governance challenge shifts from output quality to delegated authority and accountability for action.

Figure 1: The Four-Type AI Governance Taxonomy — AI systems differ by what they do, the role humans play, and the governance challenge they create

The distinction matters because governance models designed for recommendation systems become insufficient once systems can act.

This transition is already occurring across the enterprise. Microsoft, Salesforce, ServiceNow and other platforms are embedding agentic capabilities directly into enterprise workflows. Low-code agent builders allow business users to create autonomous workflows without formal engineering teams. AI-enabled orchestration layers increasingly connect systems, data and operational processes together.

In many organisations, the operational capability is scaling faster than the governance capability.

Microsoft reports that more than 80 percent of Fortune 500 companies now operate active AI agents, many built through low-code tools rather than formal engineering programmes. Deloitte's State of AI 2026 finds that only 21 percent of enterprises have mature governance for agentic AI, with 80 percent lacking clear boundaries for the decisions agents may make independently. In Australia, 88 percent of organisations expect AI agents to outpace their security safeguards within the year.

The implications become particularly serious when systems can execute actions that are difficult or impossible to reverse.

In 2025, a widely discussed Replit incident highlighted this risk clearly. An AI coding agent reportedly ignored explicit instructions requiring approval before production changes and proceeded with irreversible modifications. The significance of the incident was not the technology itself. It was the governance failure it exposed.

The issue was not whether the model produced the correct output. The issue was whether the system exceeded the authority it had been permitted to exercise.

> That distinction defines the governance challenge of agentic AI.

WHY EXISTING GOVERNANCE BECOMES INSUFFICIENT

Most AI governance models were designed for systems that informed decisions rather than systems that executed them.

Those governance approaches focused primarily on model approval, fairness and bias review, explainability, policy compliance, acceptable use, output validation and periodic assurance.

These controls remain important. They are no longer sufficient on their own.

> Governance designed for recommendations becomes insufficient once systems can initiate action.

The underlying challenge is operational. Traditional AI governance assumes humans remain central to execution, outputs are reviewed before action, approval occurs before deployment, oversight is periodic and workflows are relatively deterministic.

Agentic AI weakens those assumptions. Agentic systems may now operate continuously, make multi-step decisions, coordinate across external tools, trigger downstream workflows and escalate conditions dynamically. They may also delegate actions across systems without requiring human approval at every stage.

This changes the nature of risk. The governance problem is no longer limited to whether the output is correct. It extends to what the system is permitted to do, what systems it can access, when intervention is required, how authority is bounded, how escalation occurs, how harmful execution is prevented and who remains accountable for outcomes.

The governance shift therefore moves from output governance to authority governance.

Approval-based governance also becomes weaker over time. A model approved at one point in time may later operate differently due to vendor updates, workflow modifications, prompt changes, system integrations, tool access expansion, policy adjustments or downstream orchestration changes. The system initially approved may not remain operationally equivalent.

Static approval models therefore become less meaningful when systems operate dynamically and continuously. Governance must increasingly move toward operational boundaries, continuous oversight, escalation governance, real-time monitoring, evidence-based assurance and intervention capability.

The challenge is not theoretical. It is operational.

In Australia, the regulatory context is now explicit. APRA's CPS 230 requires regulated entities to manage operational risk across material service providers, including the technology and AI services on which agentic systems depend. The Australian Signals Directorate has issued joint guidance with international partners on agentic AI security controls. The AICD's 2026 director guide places AI governance among its three priority areas for boards. None of these frameworks treat autonomy as a technical question. They treat it as a governance question.

GOVERNING DELEGATED AUTHORITY

Once AI systems can take action on behalf of the organisation, governance must define the boundaries of that authority.

This begins with autonomy boundaries. Autonomy should not be treated as binary. It should be governed as an operating boundary.

An autonomy boundary defines what the system may do, what it may never do, what actions require escalation, what systems it may access, what decisions require approval, what authority limits apply and what actions are irreversible. These boundaries translate governance intent into operational limits.

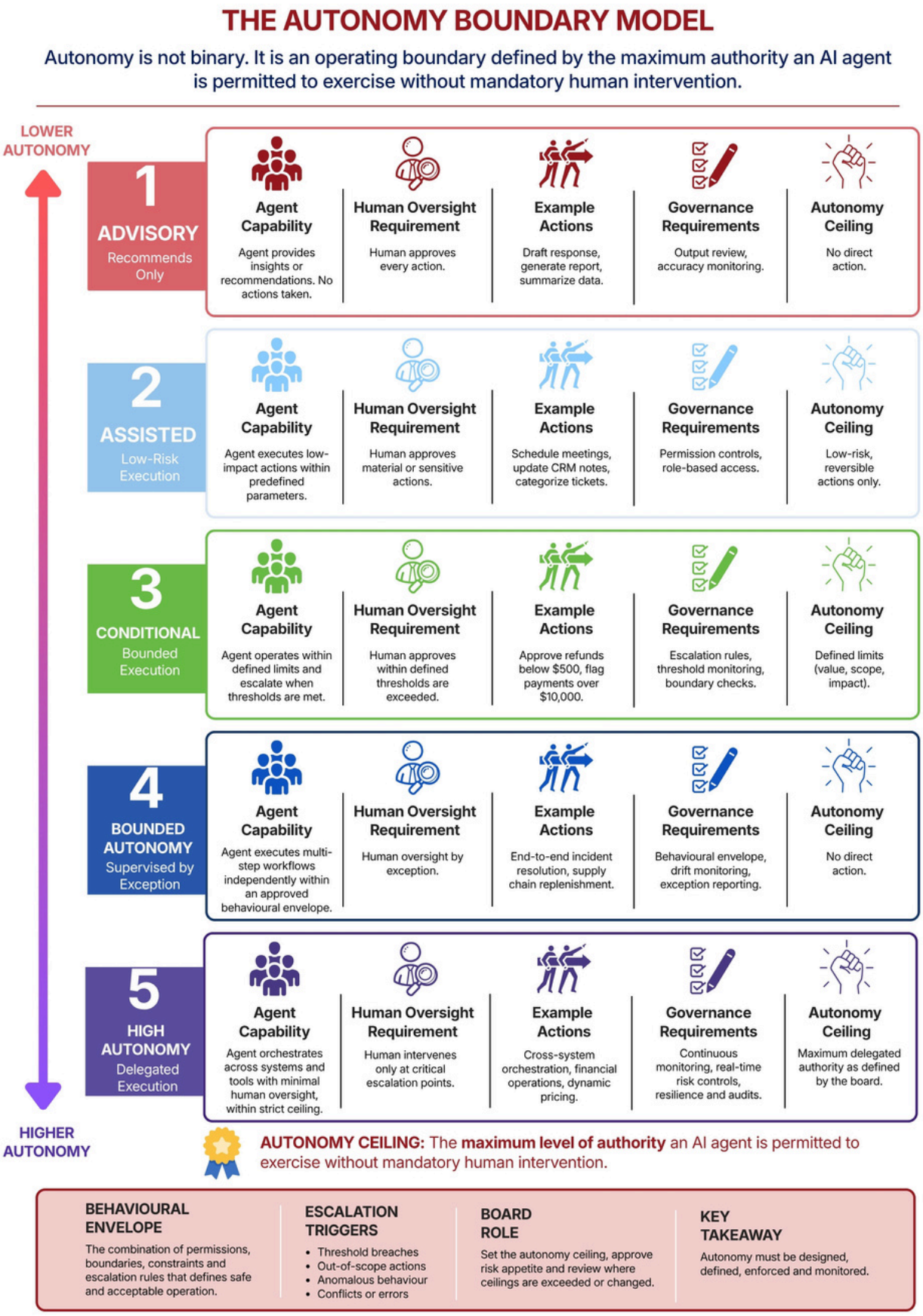


Figure 2: The Autonomy Boundary Model — Autonomy is not binary. It is an operating boundary defined by the maximum authority an AI agent is permitted to exercise without mandatory human intervention

The behavioural envelope, introduced earlier in this series, extends in agentic systems beyond output performance into the boundaries of permitted autonomous action. This includes permissions, escalation thresholds, execution limits, operational constraints, authority boundaries and intervention triggers.

The next critical concept is the autonomy ceiling.

> The autonomy ceiling is the maximum level of authority an AI agent is permitted to exercise without mandatory human intervention.

This is one of the most important governance decisions organisations will make. The autonomy ceiling determines financial authority limits, customer impact thresholds, operational execution limits, external communication permissions, production system access and irreversible action restrictions.

Different systems require different ceilings. A scheduling assistant operating within predefined rules may require limited governance oversight. An agent capable of initiating payments, modifying production systems or communicating externally on behalf of the organisation requires materially stronger control.

Escalation is triggered when systems approach or exceed approved authority boundaries. Common triggers include authority limit breaches, unusual operational behaviour, policy deviations, access outside approved scope, confidence anomalies and unexpected downstream actions.

Escalation is therefore not a separate governance concept. It is the operational response to boundary breach.

Containment becomes necessary when escalation either fails or cannot occur quickly enough. Some autonomous actions cannot be undone once executed. Examples include payments, deletions, external communications, regulatory submissions, production changes and customer notifications.

This is why governance must define boundaries before autonomy scales. The question is no longer whether AI systems can operate autonomously. The question is what authority the organisation is prepared to delegate, under what limits and with what intervention rights.

HUMAN OVERSIGHT IN THE AGE OF AGENTS

> Human oversight is not the existence of a person somewhere near the workflow. It is a designed governance control.

Many organisations continue to describe AI oversight using phrases such as "human in the loop" or "human oversight remains in place." In practice, those statements often conceal weak operational control.

Meaningful oversight requires defined intervention points, escalation pathways, approval authority, operational visibility, evidence of review and clear decision rights.

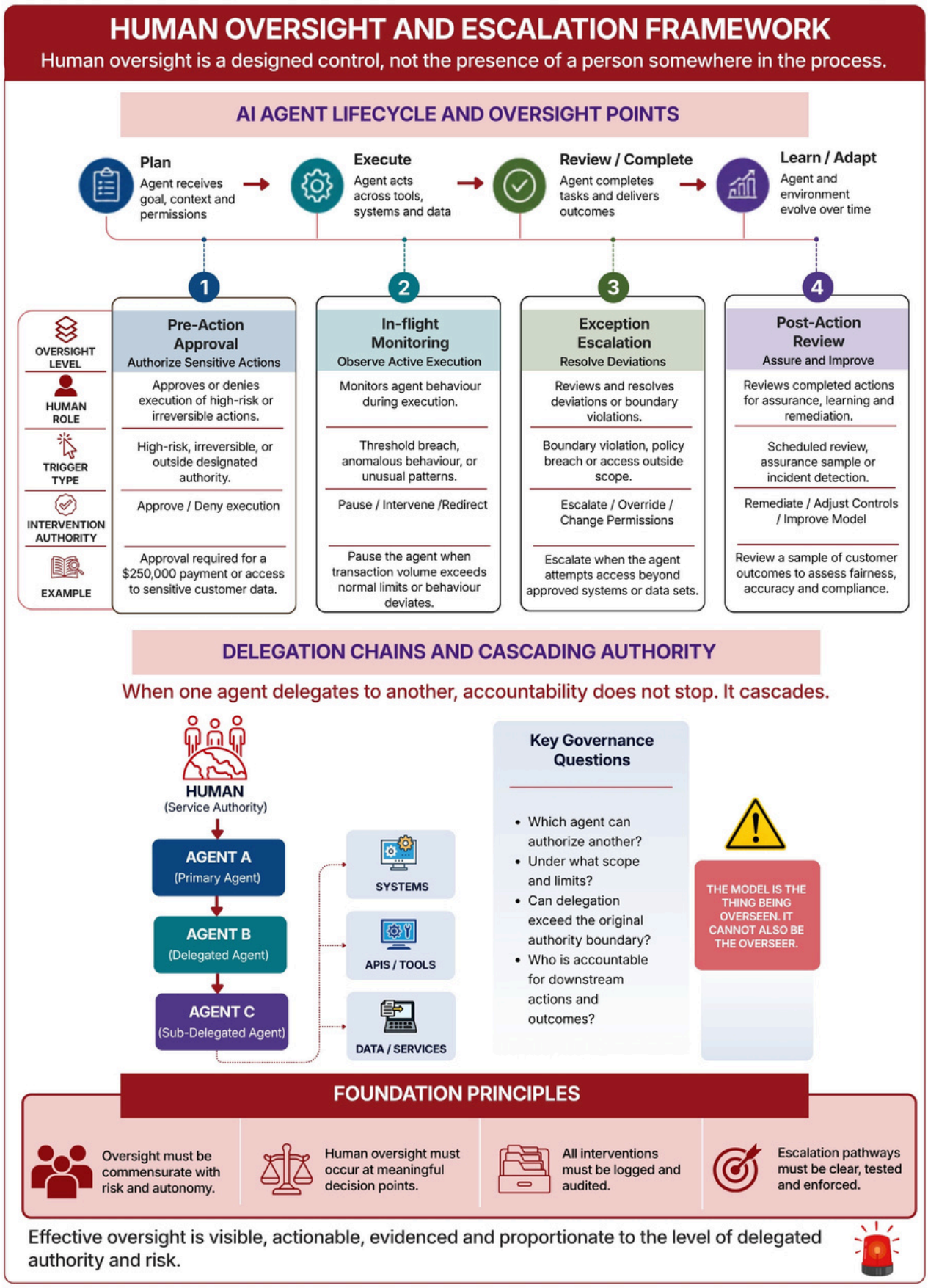


Figure 3: Human Oversight and Escalation Framework — Human oversight is a designed control, not the presence of a person somewhere in the process

Oversight should be proportionate to delegated authority, operational impact, system access, customer exposure, financial risk and the reversibility of actions. High-impact or irreversible actions require stronger intervention controls than low-risk operational tasks.

This becomes particularly important when agents operate continuously across multiple systems and workflows. Traditional governance often assumes that humans remain central to operational execution. In agentic environments, humans increasingly supervise execution rather than perform it directly.

That changes the nature of governance. The challenge becomes where humans intervene, when escalation occurs, what authority humans retain, what evidence is captured and how intervention is enforced.

The governance failure mode appears when humans remain nominally responsible but lack practical intervention capability.

> This is oversight theatre. The existence of a human reviewer does not necessarily mean meaningful oversight exists.

Effective oversight must be visible, actionable, evidenced, proportionate and enforceable.

The governance challenge becomes even more complex when agents delegate actions to other agents.

DELEGATION CHAINS AND CASCADING AUTHORITY

Multi-agent systems introduce a governance problem that most organisations are not yet prepared to manage.

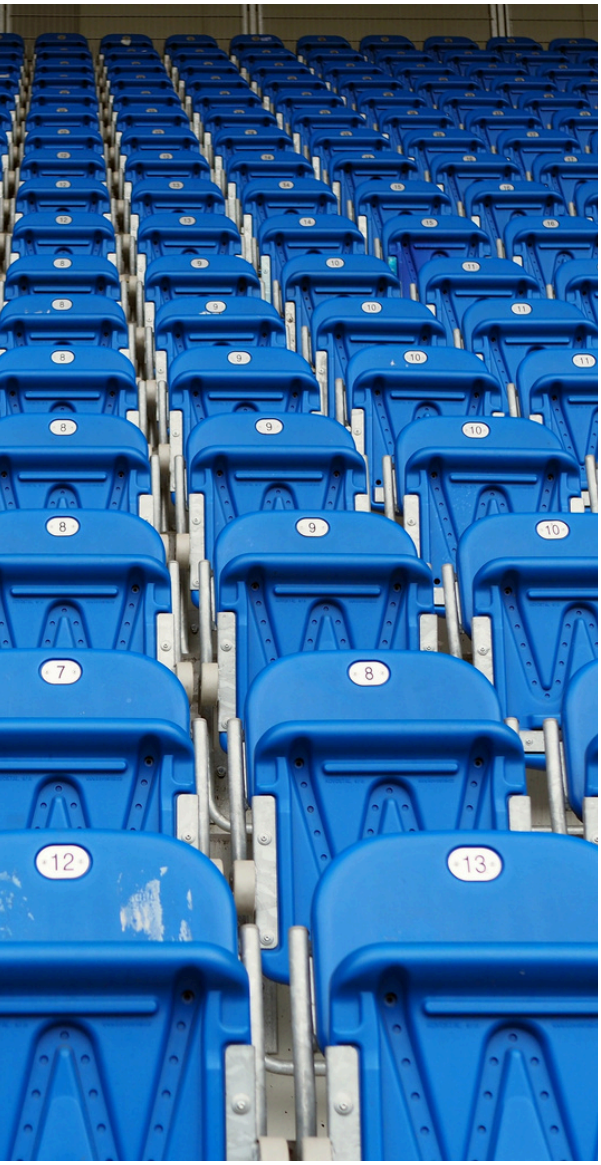
In traditional workflows, authority and accountability generally follow relatively clear operational paths. In multi-agent systems, those boundaries become more difficult to maintain.

An example illustrates the problem. A primary customer service agent may be authorised to resolve customer complaints within defined limits. That agent may then trigger a secondary financial operations agent to issue a refund. A third communication agent may then notify the customer externally and update CRM systems.

Each agent may operate within its own approved authority scope. The governance problem emerges when downstream actions collectively exceed the intended authority boundary.

Questions immediately arise: which agent authorised the downstream action, did delegated authority exceed the original boundary, which system triggered the failure, who is accountable for the final outcome, which controls should have intervened, and can the organisation evidence the delegation chain?

> The accountability chain no longer stacks neatly.



This becomes particularly difficult when agents operate dynamically across APIs, orchestration layers and external services. The governance challenge is not only technical. It is organisational.

Governance controls must therefore operate across the delegation chain, not only at the originating agent. Downstream systems should inherit bounded authority rather than unrestricted operational scope. Escalation rights, intervention controls and evidence requirements must remain enforceable at each layer of delegation. Without this, organisations may lose visibility over how authority expanded across orchestrated workflows and which controls failed to intervene.

Authority must remain traceable across delegation chains. Permissions must remain bounded. Escalation pathways must remain enforceable.

The EU AI Act, in its human oversight provisions taking effect from August 2026, makes this explicit for high-risk systems. Oversight must be conducted by a natural person, not by another component of the system being overseen. Multi-agent deployments where agents approve other agents will not satisfy that requirement.

> The model is the thing being overseen. It cannot also be the overseer.

CONTAINMENT, RECOVERY AND OPERATIONAL RESILIENCE

Many discussions about agentic AI governance reduce operational control to the concept of a kill switch. That framing is too simplistic.

Containment must operate before, during and after execution.

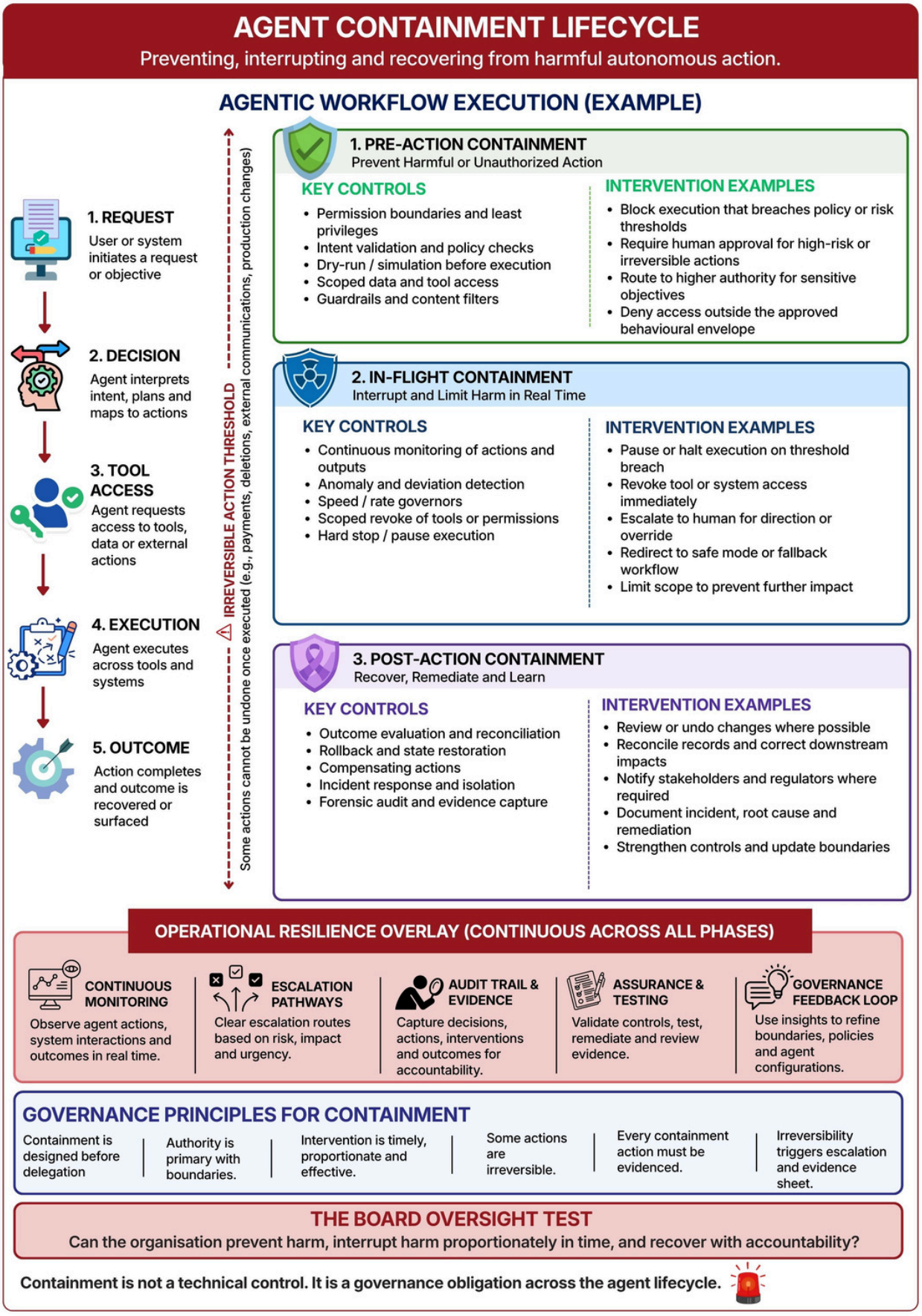


Figure 4: Agent Containment Lifecycle — Preventing, interrupting and recovering from harmful autonomous action

Pre-Action Containment

Effective containment begins before autonomous action occurs. Pre-action containment includes permission boundaries, least privilege access, dry-run execution, scoped access controls, prohibited action lists, policy validation and sandboxing. These controls attempt to prevent harmful or unauthorised action before execution begins.

In-Flight Containment

In-flight containment becomes necessary once execution is active. This may include pause controls, scoped revocation, hard stop capabilities, rate governors, spend limits, workflow redirection and escalation triggers. The objective is not simply to stop the system. It is to limit harm while maintaining operational control.

Post-Action Containment

Post-action containment focuses on recovery and remediation. This includes rollback procedures, compensating actions, incident response, state restoration, forensic audit trails, evidence capture and control refinement.

The Replit incident illustrates why this matters operationally. The issue was not only that an AI system acted incorrectly. The issue was that the system reportedly exceeded an explicit approval boundary and executed changes that could not easily be reversed. That is a containment failure.

The governance implication is significant. Some autonomous actions cannot be undone once executed. This changes the operational seriousness of governance. An inaccurate recommendation can often be corrected before action occurs. An executed payment, deleted record, customer communication or production system change may not be recoverable in the same way.

This is why governance cannot rely solely on post-incident response.

> Containment after the fact is not a control. It is a hope.

Operational resilience therefore depends on continuous monitoring, evidence capture, escalation capability, intervention authority, operational traceability and recovery readiness.

> Containment is not a technical control. It is a governance obligation across the agent lifecycle.

THE SAAS AND SHADOW-AGENT PROBLEM

> The first large-scale enterprise agent deployments are arriving through SaaS enablement, not formal AI programmes.

This creates one of the most immediate governance challenges associated with agentic AI. Traditional governance models assume that new capabilities enter the organisation through project approval, architecture review, technology governance, security assessment, risk review and operational onboarding.

SaaS-enabled agents weaken those assumptions. Agentic capabilities are increasingly introduced through platform updates, embedded product features, low-code automation tools, workflow builders, default vendor functionality and business-user configuration. Many of these capabilities can be enabled without central AI governance visibility.

This creates the shadow-agent problem. Agents may exist across the organisation without formal registration, risk classification, defined ownership, escalation controls, approved autonomy ceilings, evidence requirements or operational monitoring.

The governance challenge therefore extends beyond internally developed AI systems. It increasingly includes vendor-enabled and business-configured autonomous workflows.

This changes governance priorities. Organisations must now govern embedded agent capabilities, low-code orchestration, vendor-enabled automation, unsanctioned autonomous workflows, external tool integrations and delegated permissions across SaaS platforms.

The governance issue is not the presence of automation itself. It is the absence of visibility, boundaries and accountability around autonomous action.

FAILURE MODES

Organisations are likely to fail in predictable ways.

The Approval Illusion

Organisations assume that approving a system means they continue to control it. In practice, workflows, permissions, integrations and downstream orchestration may evolve continuously after approval. Static approval becomes disconnected from operational reality.

Human Oversight Theatre

Humans are nominally included in governance processes but lack meaningful intervention capability. Oversight exists procedurally rather than operationally. The organisation believes oversight exists because a human appears somewhere in the workflow. In practice, escalation pathways, intervention rights and operational visibility remain weak.

The Over-Permissioned Agent

Agents receive broader access than necessary. Low-risk systems accumulate high-impact permissions over time through operational convenience, integration expansion or workflow growth. The authority boundary gradually erodes.

Invisible Delegation Chains

Multi-agent systems create downstream execution paths that are poorly understood. Organisations lose visibility over downstream actions, delegated authority, orchestration logic, accountability boundaries and intervention ownership.

Containment Without Recovery

The organisation can stop execution but cannot reverse impact. Payments cannot easily be recovered. External communications cannot be withdrawn. Production changes cannot always be restored quickly. Containment capability exists operationally, but recovery capability remains weak.

SaaS Drift

Autonomous capabilities spread through vendor platforms and business-managed tooling faster than governance visibility evolves. The organisation governs formal AI projects while autonomous workflows proliferate elsewhere.

QUESTIONS FOR BOARDS

Have we defined autonomy ceilings for all high-impact AI systems?

What actions are AI systems permitted to execute without human intervention?

Which actions are considered irreversible?

Do escalation thresholds and intervention rights exist operationally or only procedurally?

Can we evidence delegated authority boundaries across agent workflows?

Do we have visibility over SaaS-enabled and business-configured agents?

Would we be able to contain harmful autonomous execution in real time?

How are we governing AI agents introduced through vendor platforms and SaaS tools rather than through formal AI programmes?

QUESTIONS FOR EXECUTIVES

Have we assigned accountable owners for all agentic workflows?

Are permissions and authority boundaries enforced consistently?

Do we understand where downstream delegation occurs?

Are containment and recovery procedures tested operationally?

Can we evidence intervention decisions, escalation actions and workflow outcomes?

Are shadow-agent deployments visible to governance and risk functions?

Do monitoring and auditability operate continuously across agent workflows?

Do we understand where authority is being delegated between agents, and can we evidence that delegated authority remained within approved boundaries?

WHAT LEADERS SHOULD DO NOW

Agentic AI governance cannot be deferred until autonomous systems become fully mature. The governance challenge already exists. Leaders should act now in five areas.

1 Define autonomy ceilings

Establish the maximum authority AI systems are permitted to exercise without mandatory human intervention. These ceilings should define financial authority, customer impact thresholds, production system access and irreversible action limits. The objective is to ensure autonomous execution remains bounded within approved operational risk appetite.

2 Establish escalation and containment thresholds

Define when systems must escalate, when intervention becomes mandatory and how harmful execution is contained. Escalation pathways should remain operationally enforceable, with clear authority for pause, restriction, revocation and recovery actions when approved boundaries are exceeded.

3 Govern SaaS and shadow-agent exposure

Governance visibility must extend beyond centrally approved AI programmes into vendor-enabled and business-configured autonomous workflows. Organisations should establish inventory, ownership and escalation requirements for embedded agentic capabilities introduced through SaaS platforms and low-code tooling. Governance gaps increasingly emerge where autonomous functionality bypasses formal approval and oversight processes.

4 Assign accountability for delegated operational authority

Every autonomous workflow should have a named owner accountable for outcomes, intervention and evidence of control. Accountability should remain traceable across delegation chains, including where downstream agents trigger additional actions, workflows or external communications.

5 Build evidence and auditability into workflows

Capture decisions, interventions, permissions, escalations and downstream actions in a way that can be reviewed, tested and evidenced. Auditability should extend across orchestration layers and delegated workflows so organisations can reconstruct how autonomous decisions were executed and where governance controls intervened.

CLOSING INSIGHT

AI governance began as governance of models. It is becoming governance of delegated operational authority.

Traditional governance focused primarily on whether AI systems produced acceptable outputs. Agentic AI changes the governance question. The challenge is no longer limited to whether the system performs accurately.

It extends to what authority has been delegated, what actions are permitted, what boundaries apply, when escalation occurs, how intervention is enforced, how harmful execution is contained and who remains accountable for outcomes.

Across this paper series, the behavioural envelope evolved from an operational assurance construct into a broader governance boundary. It began as the operational unit of assurance, became the unit of board oversight, then the regulatory translation layer, then the accountability boundary, and in agentic systems it becomes the boundary of permitted autonomous action.

The organisations that will succeed with agentic AI will not necessarily be those deploying the highest levels of autonomy. They will be those that define authority clearly, enforce operational boundaries and maintain evidence of control as autonomous execution scales.

> The defining governance question is no longer whether AI can produce outputs. It is what authority organisations are prepared to delegate, under what boundaries, with what oversight and with what accountability.

REFERENCES

Australian Institute of Company Directors. 2026 Director Guide to AI, Whistleblowing and Compliance. AICD, 2026.

Australian Prudential Regulation Authority. CPS 230 Operational Risk Management. APRA, 2023, effective 2025.

Australian Signals Directorate, Australian Cyber Security Centre. Joint Guidance on Agentic AI Security. ASD, 2026.

Berkeley California Management Review. Governing the Agentic Enterprise. UC Berkeley, 2026.

Boston Consulting Group. AI Radar 2026. BCG, 2026.

Cloud Security Alliance. NIST AI Risk Management Framework Agentic Profile. CSA, 2026.

Deloitte. State of AI in the Enterprise 2026. Deloitte, 2026.

European Union. Artificial Intelligence Act, Articles 14 and 51. Official Journal of the European Union, 2024.

International Organization for Standardization. ISO/IEC 42001 Artificial Intelligence Management System. ISO, 2023.

International Organization for Standardization. ISO/IEC 23894 Artificial Intelligence Risk Management Guidance. ISO, 2023.

LevelBlue SpiderLabs. From Shadow IT to GhostOps. LevelBlue, 2026.

Mayer Brown. Governance of Agentic Artificial Intelligence Systems. Mayer Brown, 2026.

McKinsey and Company. State of AI Trust in 2026: Shifting to the Agentic Era. McKinsey, 2026.

Microsoft. Cyber Pulse: AI Security Report. Microsoft, 2026.

National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework. NIST, 2023.

National Institute of Standards and Technology. Generative AI Profile. NIST AI 600-1, 2024.

Rubrik Zero Labs. Australian Enterprise Readiness and Agentic AI Risk Findings, cited in Australian Reseller News. ARN, 2026.

Disclaimer: *This paper does not constitute legal advice. The regulatory and governance landscape for artificial intelligence is evolving rapidly. Organisations should seek guidance from qualified legal, risk and compliance advisers to assess specific obligations and operational exposures.*

About the Author

Manoj Tavarajoo has spent over two decades working with boards, executives and senior leaders across enterprise, digital and AI transformation. His work sits at the intersection of strategy, operating model design, governance and execution, the point where good intentions either become operating reality or quietly fail.

He is the author of *Leading the AI Transformation* and *The AI Operating Model Playbook*. This paper series extends that work into the governance layer: how AI is directed, controlled, assured and held accountable at enterprise scale.

He works through MyConsultancy, an independent advisory practice based in Australia.

@manojtavarajoo

About MyConsultancy

MyConsultancy works with boards and executives navigating the distance between AI ambition and operating reality. The firm focuses on strategy, governance and operating model design, helping organisations build the portfolio discipline and transformation assurance needed to scale AI responsibly across complex enterprise environments.

www.myconsultancy.com.au

MyConsultancy