pwc

The AI trust dividend

Three Governance Shifts to Build Trust in Agentic AI Autonomy

September 2026

Table of contents

Introduction

AI agents[1] require governance that combines explicit accountability, operational oversight, and controls that scale as agent use multiplies.

These evolving tools introduce agency and autonomy[2] but lack true intelligence or accountability, which raises questions about trust and reliability in critical activities — and trust among stakeholders is what turns AI into a strategic advantage. Trust gives stakeholders confidence in business decisions, products, ethics, behaviours, and governance. In contrast, a failure to deliver trust can result in loss of company value and brand loyalty, as well as risk legal or regulatory action.

Trust in autonomous AI agents will arise from governance and controls that evolve beyond traditional approaches and methods used in early AI use case experimentation. Once trust is established, it produces a dividend. PwC's AI Performance study shows that organisations can deploy AI faster and capture dramatically higher returns than their peers.

In this third article in our AI trust dividend series, we identify the governance focus, operating model, and accountability shifts leaders need to make to scale agentic AI safely, compliantly, and with confidence.

[1] AI Agents: AI systems that can interact with their environment, receive information, and undertake self-directed actions in service of a larger, externally specified goal.

[2] Autonomous AI: The capability of a system to select between multiple possible action sequences to achieve the system's goals, based on the current situation and internally defined criteria. Current AI systems do not yet exhibit autonomy in the fullest sense of this term. While emerging agentic systems can independently plan, select tools and execute sequences of actions, their goals, permissions, operating boundaries and available actions remain externally defined. Their apparent autonomy is therefore better understood as "bounded or delegated autonomy" within a human-designed operating envelope, rather than independent determination of objectives or unconstrained action.

01

What's different about agentic AI?

Five characteristics make agentic AI different than its predecessors.

- Depending on the use case design, agentic AI can reason over goals and context, rather than simply following predefined rules.
- It can make decisions between possible courses of action based on objectives, context, and constraints.
- It can take actions autonomously across systems, tools, and processes.
- Its behaviour is non-deterministic, drawing on insights from patterns, context, and prior interactions, and producing outcomes that cannot always be predicted with certainty.
- When combined with memory and feedback, it can learn and adapt, causing behaviour to evolve over time.

This combination of traits creates significant opportunities for innovation and transformation. AI leaders are leaning into these opportunities. Our AI Performance study shows that AI leaders, defined as the 20% of organisations that are capturing 74% of AI's economic value, are twice as likely to use AI that operates autonomously than their peers. However, opportunity and risk go hand in hand. AI agents demand new approaches to governance, risk management, oversight, and scalable controls to ensure safety, compliance, and reliability.

02

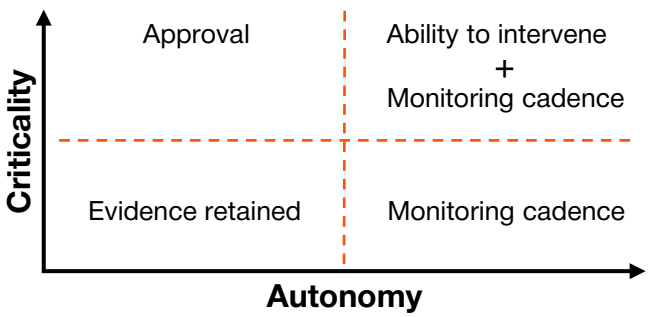
Governance implications

As the rate of AI adoption accelerates beyond many organisations' governance maturity, they are pushed to choose between slowing AI implementation or improving governance responsiveness. With competitive pressure building, the race is on to more closely match governance capability to competitive pressures in order to accelerate AI innovation and enable AI adoption at scale.

Governance needs to evolve from a focus on controlling AI technology deployments to also governing operational AI decision-making and actions. In this evolution, governance arrangements should be influenced by two factors: the agent's level of autonomy and the criticality of the role and decisions the agent is performing. Criticality often increases as agents interact directly in the organisation's value chain.

Thinking of governance in the context of autonomy and criticality helps to calibrate the proper approach. For example, the control posture for agents with low criticality and autonomy could be to retain evidence. Those with higher criticality but low autonomy may need approvals.

High autonomy but low criticality agents could require a monitoring cadence and agents with high autonomy and criticality may need the ability to intervene based on a continuous monitoring cadence. It is for these agents operating at the edge of the two spectrums where continuous, technology-enabled oversight is essential to match the pace and scale of agentic operations.



In practice, governance of agentic AI must address both the underlying AI system and the business activity it performs. Traditional controls over models, data, testing, and deployment remain important, but they are no longer sufficient. For each agent, governance must now address two distinct questions. Getting these two aspects right will provide much higher confidence in the outputs generated by the AI system.

1. Is the agent doing things the right way (execution) in areas such as performance, data protection, security, monitoring, and operational compliance?
2. Is the agent doing the right things (intent), meaning is it pursuing the right objectives within its delegated authority and risk appetite? Every agent needs clear limits: what it can decide alone, and when it must escalate to a human.

The true source of intent lies with human designers, builders, and business owners. Governance must therefore examine how intent is encoded into the system and how ongoing evaluation of its adherence is performed.

In addition, governing the combined system of agents working together becomes materially more important, including its operational behaviour, orchestration, permissions, tools, memory, decision rights, actions, escalation, and human oversight.

03

Three necessary shifts

Agentic AI governance requires leveraging existing and new capabilities in a highly coordinated, automated, and continuous manner. As organisations revisit how decisions are made and by whom, or what in the case of agentic AI, there are three primary shifts to consider.

Shift 1: Human accountability at scale

Human-in-the-loop (HITL) controls can provide an important early-stage mechanism for learning how agents behave in real-world environments, validate outcomes, and retain accountability.

However, human-in-the-loop supervision is no longer sustainable where agents increase in number, pace, and complexity. Here, we see an evolution towards more scalable human accountability with supervision increasingly embedded and engineered. What does agent supervision and accountability look like in practice when the ratio of human to agent could be one to hundreds? As AI agents undertake more monitoring and operational oversight, humans remain accountable for outcomes, decisions, and risk appetite. Every agent needs a business owner who is accountable for the agent's decisions and actions.

The Board should oversee whether the organisation has effective governance for agentic AI, while executive and functional management should establish clear ownership, escalation rights, and decision accountability, rather than directly supervising every agent activity.

As a practical step to facilitate human accountability, some organisations are building new user interfaces to present information (inputs, outputs) to humans such that they can assess whether an agent has made appropriate recommendations and decisions. This includes working through how to best present the path taken by an agent in a way which is consumable, understandable, and connected to the overall purpose of an agentic workflow or a group of agents — thus enabling a control decision to be made. A minimum decision record often includes the delegated objective, action taken, evidence or rationale, any exception flags, the accountable owner, and escalation outcome. Over time, trust in agentic operations grows and this AI trust dividend allows for scaled adoption within risk appetite.

Shift 2: Operating model adaptation

Governance must keep pace with the rate of innovation and the fast-growing volume of agentic activity. Even for those who decide to go slow, areas such as cyber security need to consider advanced agentic AI use and familiarity to remain adequately protected from adversarial threats.

An effective operating model ensures that autonomous systems remain observable, measurable, and auditable throughout their lifecycle. Continuous observability enables oversight in dynamic AI environments, while rapid sensing, decision-making, and intervention capabilities help organisations manage risks in real time.

In concept, organisations should establish an operating model that recognises agents as digital workforce counterparts to employees, utilising similar governance principles. Every agent should have a verified identity, a clearly defined role, task-specific access rights, auditable activity logs, and well-defined boundaries for autonomous decision-making. As agents gain greater autonomy, corresponding levels of human oversight should be strengthened, particularly for actions involving customers, employees, sensitive data, financial outcomes, or legal and regulatory obligations.

A mature model mirrors the principles used to govern human workforces, granting authority based on role and responsibility (in the context of process criticality and autonomy), while recognising that agents require a higher degree of control.

Shift 3: Ongoing governance oversight

With agentic AI, governance emphasis shifts from predominantly pre-production controls to a balance between design controls and continuous operational monitoring, evaluation, and oversight. Governance post launch should retain sufficient, risk-proportionate evidence of material actions, decisions, tool use, and exceptions to enable investigation and intervention.

Post-launch observability can encompass understanding the agent's full runtime behaviour including what inputs it received, how it reasoned or orchestrated, which tools it invoked, what actions it took, what permissions it exercised, what state changed, and what outcome resulted. In some scenarios, full action logging such as this is most appropriate. However, in others this may be unfeasible, privacy-sensitive, or disproportionately costly.

Control methods such as LLM-as-a-judge technology (using AI to test and evaluate AI) can help in pre-production testing or ongoing observability with sufficient controls over their own calibration, bias, reproducibility, and false-negative risk.

The shift to continuous oversight also requires a parallel shift in budgeting from project-based to business-as-usual funding and resourcing for agentic governance oversight. Consideration should be given to the total cost of agents, inclusive of developing, testing, managing, and optimising each agent over its lifetime.

04

Risks and controls in practice

In application, the risk categories agentic AI governance must address are familiar — it is risk exposure and control design that change. For example, data, legal and compliance, and process risks are often well-understood categories, but agentic execution changes their scale, speed, and control points. Infrastructure risk requires new thought, particularly as it relates to using a foundational model provider, agentic harnesses, and the run-time environment that underpins every agentic operation.

Model risk (such as hallucinations) and model orchestration across agentic platforms are often new territories as is use risk (such as human over-reliance on AI outputs). We find these are typically areas of greater challenge for organisations given the technology's rapid evolution.

In addition to risk exposure and controls considerations, four process priorities repeatedly emerge from our work helping organisations confidently scale agentic AI:

- **Observation and evaluation are critical:** If such monitoring is not possible on a process or data quality output controls basis, agent telemetry or model orchestration is required to maintain confidence that agents continue to operate within risk and performance tolerances. Per our operating model analogy, deploying agents is a little like hiring humans, with ongoing objective alignment and performance evaluation key to their ongoing success.
- **Agent control blueprints quickly become a necessity as agent use multiples:** By creating control blueprints based on reusable patterns of agents sharing the same criticality, risk, and autonomy levels, and then making them accessible on the organisation's agent platforms, new agents can be deployed faster from both operational and compliance perspectives. Increasingly these blueprints don't just guide how agents operate, but also how they are supervised, governed, and optimised over time.

- **Done well, an agent inventory expands to performance and value monitoring:** Maintaining an agent inventory is necessary for compliance and oversight. It also lessens the risks of agent proliferation which, over time, adds cost, complexity, and risk.To inventory effectively, information flows need to be consolidated into one cockpit which then facilitates the integration of performance and value monitoring.As this occurs, the AI governance function can evolve from a compliance focus into a governance and value capability, helping the organisation understand not only whether agents are operating safely, but whether the remit granted to them is actually delivering value.
- **Business risk becomes the most critical lens:** As agentic AI use accelerates and AI governance matures, risks and controls discussions previously focused on compliance risk shift to business risk. Business risk is firmly centred in the business with the support of strong governance. Business confidence in agent autonomy is dependent on governance maturity.

Of utmost importance is that organisations have the right foundations in place but maintain flexibility for the future as the need for new capabilities arises.

05

Building trusted autonomy

The emergence of autonomous AI agents requires organisations to move beyond viewing AI as merely models, data, or code. Agentic systems introduce new forms of autonomy into organisations, challenging existing governance boundaries and assumptions. They can transform how work gets done, but granting autonomy requires confidence.

That confidence comes from orchestrated governance and a modern control environment that moves beyond human review and relies on continuous, technology-enabled oversight with human accountability.

06

Actions for leaders

Here are three questions to ask within your organisation to proceed with confidence as you scale your use of agentic AI:

- Can we evidence ownership, authority, and escalation for every agent across the portfolio?
- Can our operating model detect, decide, and intervene at machine speed where risk demands it?
- Do autonomy and criticality determine the control intensity, evidence, and review cadence for each agent class?

Organisations that are actively addressing these governance shifts are building the responsiveness and oversight needed to trust AI outcomes. That trust strengthens their ability to deploy AI more quickly and scale it more confidently. As a result, they are better positioned to capture the innovation and transformation potential that a portfolio of autonomous agents can offer.

This is the third article in ‘The AI trust dividend’ series, where we continue to go deeper on the challenges business leaders need to address at different points in their AI maturity to unlock its full potential.

Authors and contributors

Gayan Benedict
Partner, Advisory, PwC Australia
gayan.benedict@au.pwc.com

Hendrik Reese
Partner, Advisory, PwC Germany
hendrik.reese@pwc.com

Micah Richard
Partner, Assurance, PwC United States
micah.richard@pwc.com

Mo Meskarian
Director, Advisory, PwC United Kingdom
m.meskarian@pwc.com

Thank you

Access the online version