Governance Institute of Australia

Another Good Decision

WHITE PAPER ON

Governance in the age of agentic AI

Agentic AI Oversight, Governance and Risk Management. The future of work for business leaders.

Proudly sponsored by: Diligent MALLESONS seek ethicai

Governance Institute of Australia | AI Governance Expert Panel

Contents

About Governance Institute of Australia

A national membership association, Governance Institute of Australia advocates for a community of governance and risk management professionals, equipping over 7,500 members with the tools to drive better governance within their organisation.

We tailor our resources for members in the listed, unlisted and not-for-profit sectors, and ensure our member's voice is heard loudly.

As the only Australian provider of chartered governance accreditation, we offer a range of short courses, certificates and postgraduate study to help further the knowledge and education of the fast-growing governance and risk management profession.

We run a strong program of thought leadership, research projects and news publications and draw upon our membership of the Chartered Governance Institute to monitor emerging global trends and challenges to ensure our members are prepared.

Our members know that governance is at the core of every organisation — and in these tumultuous times, that good governance is more important than ever before.

Contents

Acknowledgements

Special thanks to Daniel Popovski, Senior Advisor, Policy and Advocacy, Governance Institute of Australia, Bryony Evans, Partner, Mallesons, Lim Cheng, Partner, Mallesons, Luke Pallaras, Practice Development Senior Associate, Mallesons, Fernando Mourao, Head of Responsible AI, Seek, Kara Bombell, Co-founder EthicAI, Katriel Healy, co-founder EthicAI, and Marc Cheong Deputy Director and Digital Ethics Advisor with CAIDE, as well as members and Chair, Kylie Dalton, of the AI Governance Expert Advisory Panel for their valuable insights and editorial contributions to this document.

About the Sponsors

Governance Institute of Australia

Diligent

Diligent is the AI leader in governance, risk and compliance (GRC) SaaS solutions, helping more than 1 million users and 700,000 board members to clarify risk and elevate governance. The Diligent One Platform gives practitioners, the C-Suite and the board a consolidated view of their entire GRC practice so they can more effectively manage risk, build greater resilience and make better decisions, faster. Learn more at diligent.com

MALLESONS

The top-tier independent law firm from Australia, a full-service firm trusted on the most complex and consequential matters across our region and around the world. Our independence gives us the flexibility to combine our top tier full-service capability with the best advisers in every market. We have 200 partners and over 1,200 lawyers who are locally qualified and globally experienced.

The quality of our capability is reflected in our market standing – e.g. for 8 consecutive years, we have had more Band 1 practices and Band 1 individuals in Chambers and Legal 500 than any other firm. We're here to be the firm clients and communities trust most with their future in a world that never stands still. To have an outsized impact. To lead the market. To make a difference to be proud of. To create careers that can go everywhere.

seek

SEEK is a leading Asia-Pacific employment marketplace, connecting candidates and hirers across eight markets through AI-powered matching at scale. With over 25 years of innovation, SEEK sets the standard for Responsible AI in recruitment, building hiring experiences that are fair, transparent, safe, accountable, and human-centred.

EthicAI delivers AI for Human Dignity. We help organisations design, govern and implement AI that strengthens performance without compromising people. Through education, applied consulting and practical tools, we embed defensible decision making and workforce readiness so AI adoption is commercially sharp, ethically grounded and built to last.

The Centre for Artificial Intelligence and Digital Ethics (CAIDE) facilitates cross-disciplinary research, teaching, and leadership on the ethical, technical, regulatory and legal issues relating to Artificial Intelligence (AI) and digital technologies.

CAIDE aims to build policy and regulatory expertise at the University of Melbourne and across the wider community.

CAIDE is founded as a collaboration between the Faculty of Engineering and Information Technology and Melbourne Law School, with member faculties Arts, Education, and MDHS, with support from the University of Melbourne.

Governance Institute's AI Governance Expert Advisory Panel

The AI Governance Expert Advisory Panel (AI Advisory Panel) is an initiative of the Governance Institute of Australia. The AI Advisory Panel's mission is to collaborate across business, community, and special interest groups to help inform best practice governance of AI across Australian workplaces. The AI Advisory Panel is intended to build the requisite skills and capabilities of governance, risk and other business professionals with an interest in AI governance and deployment.

CEO Foreword

The future of work is evolving rapidly. It was only a few short years ago when we were introduced to the powers of generative AI creating excitement over new ways of working. Fast forward to 2026, and AI technologies have rapidly evolved into autonomous decision makers that have developed capabilities to learn, reason and execute decisions, exhibiting human like agency acting on our behalf. The opportunities of this powerful technology are truly limitless, evolving how we govern and manage organisations and delegate responsibilities and execute organisational strategies.

However, this new wave of AI technologies does not come without its own risks. Rushed deployment because of fears of missing out or acting quickly without contemplating the technology's technical limitations, the ethical dilemmas it raises, as well as existing and emerging legal rules and regulatory frameworks and responsibilities is almost guaranteed to lead to failed pilot programs and poorly executed business cases.

The threat of agentic AI projects being cancelled due to escalating costs and unclear business value propositions is a real prospect facing Australian organisations. This is driven by a lack of awareness into the evolving nature of risk management, governance and oversight of non-human identities in the workplace. Agent AI delegation is now critical in orchestrating a new digital workforce of agents that have the potential to drive positive changes in the workplace spilling into economic and societal value.

The Paper offers insights into the practical questions, insights and steps governance professionals and business leaders should be taking on their journey of introducing AI agents in the workplace.

We have worked closely with our sponsors, Diligent, Mallesons, Seek, EthicAI, University of Melbourne's Centre for AI and Digital Ethics and the Governance Institute's AI Governance Expert Advisory Panel to offer you a comprehensive toolkit into reimagining governance in the age of agentic AI.

I truly believe that together we can help create a better society and economy by improving the way decisions are made everywhere.

Katrina Horrobin
CEO
Governance Institute of Australia

Executive summary

Agentic AI is a powerful technology with rapidly advanced capabilities in automation independent of direct human supervision. The fast pace and autonomous nature of agentic AI systems create the potential for organisations to produce significant increases in productivity and efficiency-enhancing improvements in the workplace. The technology can execute repetitive, labour-intensive activities and execute complex workflows with limited human oversight, augmenting the way we work.

The potential productivity gains of AI are well-documented however in the case of agentic AI systems that are becoming increasingly autonomous, critical questions arise about quality, integrity and veracity wherever non-human decision making occurs.

Sound governance and risk management frameworks play a critical role in the effective governance of agentic AI technologies. Governance professionals require a solid grounding in ethical and human-centred approaches to effectively deploy and de-risk the powerful capabilities of the technology. Risk mitigation and harm minimisation are necessary foundations in the development of defensible governance frameworks for its effective rollout.

The paper is in several parts designed to build the reader's understanding. **Part 1 aims to equip the reader with a baseline standard of technical knowledge.** A general working knowledge of key technical terms and how agentic AI technologies operate is a necessary starting point for governance professionals to effectively manage and govern the technology.

Part 2 extends on these technical underpinnings to discuss the critical and emerging ethical dilemmas impacting the way in which the technology is adopted and deployed in real life settings. Contemplating ethical dilemmas in contemporary workplace settings is a core skill in effectively managing how an increasingly powerful and autonomous technology may impact on the organisation's operating environment, including its potential impact to workers, customers, suppliers and the broader community. Taking an ethical principles-based approach to deployment allow directors to develop a defensible governance strategy when deploying agentic AI systems.

Part 3 develops critical awareness of the regulatory and legal landscape associated with the use and deployment of the agentic AI technologies. This chapter aims to provide readers with sufficient confidence in understanding how existing laws may apply to the use of agentic AI technologies in workplace settings.

Part 4 provides insights into what senior officers and company directors may consider when developing an Agentic Governance and Oversight Framework. It informs readers by raising critical awareness of existing and emerging legal duties that apply, and the sorts of questions executive teams and company directors should contemplate on their Agentic AI deployment journey.

Part 5 outlines how directors can demonstrate that an AI agent's decisions and actions were reasonable, lawful, and aligned with organisational intent as well as how directors can test and verify the integrity of agentic AI systems. It also outlines questions the board be asking about their organisation's AI governance.

Part 6 builds on the technical, ethical, legal and regulatory foundations, applying governance fundamentals to offer practical use-case on how to govern AI agents as a digital workforce effectively. The future of work appears as an orchestration of multi-agent AI systems as an effective method of managing agentic workflows and outcomes, challenging traditional management practices and offering insights into a new way of working.

What is the current scale of adoption?

A recent study conducted by McKinsey across Asia, Europe and North America, finds that among the 200 C-suite executives surveyed across a wide range of enterprises, **more than 80 percent report that they are already running pilots on agentic AI**, with some progressing to scaled deployments.[1] According to McKinsey and Co, 'organisations are cautiously optimistic that agentic AI will deliver top-and bottom-line growth that Gen AI has, to date, struggled to achieve'.[2]

An international study found that **60 percent of Australian respondents say they are already deploying agentic AI technology**.[3] According to the same study **over 90 percent are using or making plans to adopt agentic AI within the next six months.**[4] This surge in investment is driven by the perceived potential productivity and efficiency uplift, with AI agents increasingly facilitating a redesign of how work is executed and coordinated across teams and systems.[5]

The value proposition of agentic AI: Is it all hype?

According to research undertaken by Gartner, **over 40 percent of agentic AI projects will be cancelled by the end of 2027 due to escalating costs, unclear business value propositions and inadequate risk and governance frameworks.**[6]

Early-stage experimentation is often driven by hype, a fear of missing out, leading to misapplied and uncoordinated agents across organisations and data governance gaps. The failure to effectively coordinate and govern agentic systems leads to unrealised returns on investment, heightened cost, risk and complexity in scaling and poor defensibility and traceability of agentic AI decisions.[7]

The development and implementation of an agentic governance and oversight framework becomes the essential starting point for the effective deployment of agentic AI systems in the workplace. Organisations that are acting to deploy agents before contemplating the governance, risk and potential compliance and legal obligations risk failing to realise the long-term, tangible benefits of this powerful technology.

---

[1] https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/reimagining-the-value-proposition-of-tech-services-for-agentic-ai

[2] https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/reimagining-the-value-proposition-of-tech-services-for-agentic-ai

[3] https://www.cyberdaily.au/tech/11928-more-than-half-of-australian-businesses-are-already-using-agentic-ai

[4] https://cfotech.com.au/story/australian-firms-to-boost-agentic-ai-investment-to-usd-12-3-billion

[5] https://cfotech.com.au/story/australian-firms-to-boost-agentic-ai-investment-to-usd-12-3-billion

[6] https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

[7] https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

Technical foundations

What are AI agents?

AI agents can independently set goals, make decisions and act with limited human oversight, integrating with other software systems to complete tasks independently or within minimal human supervision. **An AI agent is not just software that automates predefined tasks or interacts with LLMs on determined workflows, it can embody human-like agency by reasoning, learning and acting autonomously with purpose.**[8] Agents acquire knowledge and adapt to dynamic environments, they can predict and anticipate outcomes, reason and make context-aware decisions to achieve specific goals.

How do AI agents operate?

The two broad types of agents are commonly defined as reactive and cognitive agents. Reactive agents, respond to stimuli without planning, co-ordinating or setting goals.[9] Cognitive agents on the other hand are designed to perceive their environment, reason about it and make decisions based on objectives. A rational agent is a type of cognitive agent that acts to achieve the best outcome or, when there is uncertainty, the best expected outcome.[10]

Frameworks used to categorise agents may be distinguished by their logical structure. Capability driven agents sit along a spectrum of autonomy and intelligence, from simple reactive agents that react instantly to present moment stimuli, to learning agents, which improve their own performance over time. Other sub-types may include goal-based agents and utility-based agents that calculate the most desirable or efficient outcome.[11]

Architecture-driven agents are categorised by their internal logic and they may be distinguished between logic-based agents that rely on formal deduction, reactive agents that draw on stimulus-response pathways and cognitive agents that leverage goal-directed reasoning using internal models, such as Belief-Desire-Intention (BDI) architecture that manage 'mental states' representing an agent's beliefs, goals and commitments.[12] Layered architectures may stack different levels of reasoning from reflexes to high-level planning through coordinated systems.[13]

[8] https://xmpro.com/agentic-ai-for-industry-cutting-through-the-hype-to-unlock-real-industrial-value/#:~:text=Avoiding%20%22Agent%20Washing%22,it%20provide%20unique%2C%20transformational%20value?

[9] https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf

[10] https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf

[11] https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf

[12] https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf

[13] https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf

Action

Understand the technical characteristics and different types of AI agents, including how they make decisions, how they are trained to operate and their inherent capabilities when executing autonomous activities.

How are AI agents being used?

AI agents are becoming more commonly used in the workplace. The autonomous nature of AI agents raises critical questions around how organisations operate, where and how decision-making powers are delegated and exercised, how people are managing the technology and the ways in which organisational goals are overseen.

True Ai agents have much greater autonomy over how they achieve their goals and can engage in far more complex behaviour.[14] **The ability of agents to dynamically plan their next action and the open-ended nature of their process, on the one hand, makes them more flexible but, on the other hand, reduces predictability and increases the potential for errors.**[15] It is critical that every organisation embedding AI agents has a governance oversight framework. This may start with a risk assessment to evaluate the cost, benefits and long-term value the technology may deliver to the organisation.

To deliver real world business value an effective agentic AI strategy requires an understanding and systemic assessment of risks. AI agents have many different uses; from fraud detection and personalised financial advice to customer interactions and data decision support. AI agents can leverage Application Programming Interface (APIs) to communicate with other agents and humans, receive and send money, and access and interact with the internet.[16] AI agents can execute multi-step plans, use external tools and interact with digital environments to function as powerful components within larger workflows.[17] AI agents are transforming workplace environments as autonomous systems collaborate with humans to orchestrate complex workflows and unlock productivity.[18]

[14] https://www.mallesons.com/au/en/insights/latest-thinking/agentic-ai-rogue-agents-real-liability.html?redirect
[15] https://www.mallesons.com/au/en/insights/latest-thinking/agentic-ai-rogue-agents-real-liability.html?redirect
[16] https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained
[17] https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained
[18] https://www.deloitte.com/au/en/services/consulting/perspectives/human-agentic-workforce.html

Action

Consider the value proposition, risk profile and tangible benefits of agentic AI systems to orchestrate complex workflows.

What is agentic AI

Agentic AI systems can run full workflows on their own, working with people rather than being closely supervised at every step.

Agentic AI systems are semi- or fully autonomous with the ability to perceive, reason and act on their own in a way a human would. While AI agents and agentic AI share foundational characteristics, agentic AI places stronger emphasis on co-ordination among multiple agents, task decomposition and delegation, sustained operation over time, and operation in more complex and less predictable environments with limited human oversight.[19]

What are some major risks of using agentic AI systems?

Studies have found **deployment of agentic AI remains fragmented**, with 50% of agents **still operating in silos rather than as part of a multi-agent system**, driving disconnected workflows, duplicated automations, and a higher risk of **'shadow AI'** being adopted without approval or oversight.[20] **Fragmentation carries extensive risks** including inefficient operations and wasted resources, fragmented data and poor decision-making, increased risk of security vulnerabilities, cascading hallucinations and uncontrolled autonomy, as well as limited innovation and growth, higher propensity of failure to effective transformation and stifled creativity.

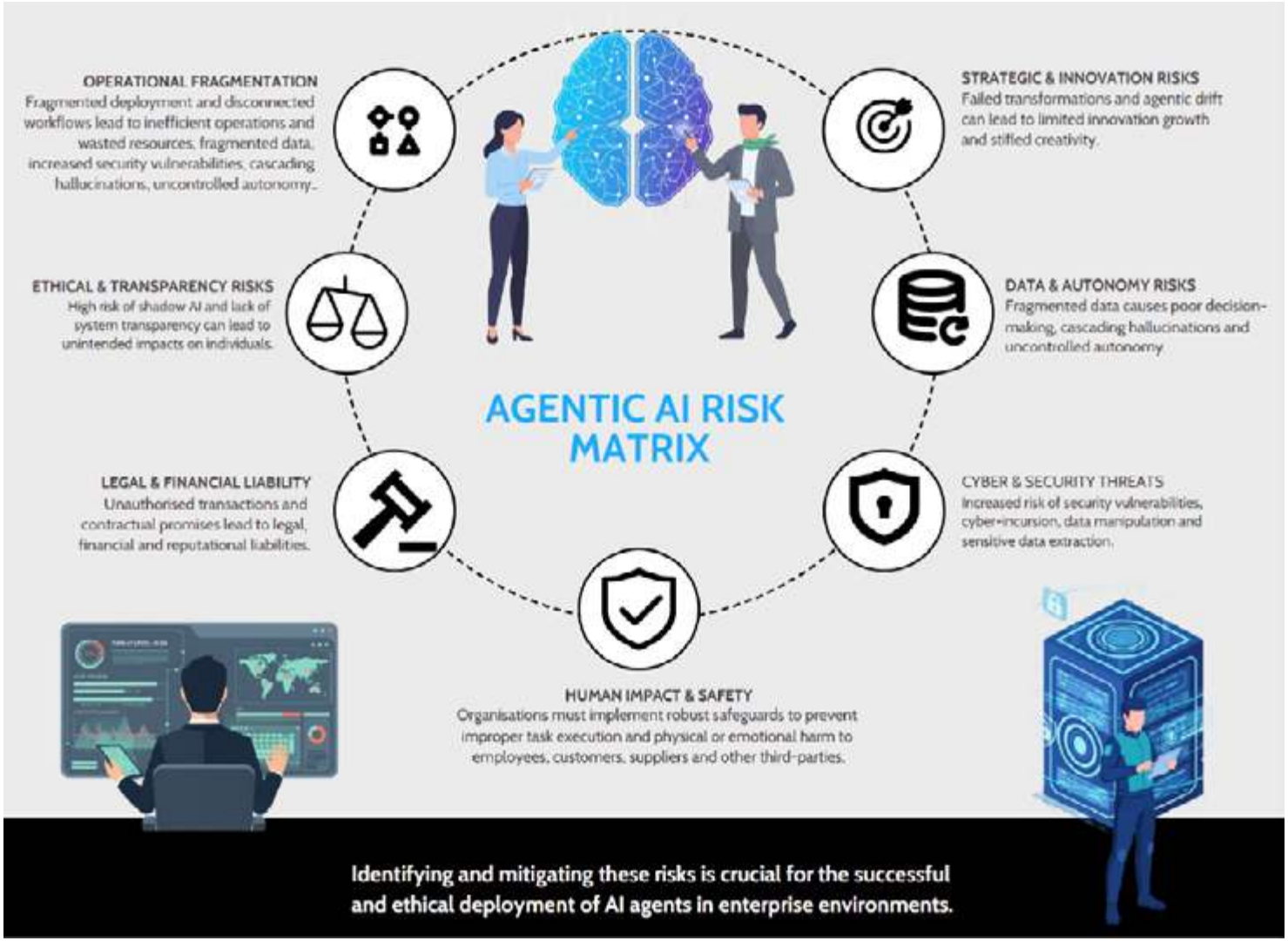
The potential for autonomous agents to **alter or delete data due to poor judgment or improper task execution** is a key risk. **Data manipulation** may also be driven by malicious actors through cyber incursions that are more easily accessed through fragmented agentic systems. A siloed agent may be convinced to **extract sensitive data or assist in the process of a cyber-attack** reinforcing the need for comprehensive cyber and data security safeguards and integration. A 'lone agent' operating within a siloed environment has the potential to compromise cyber security systems.

External communications created or sent on behalf of a company generated by AI agent may create **potential legal, reputational and financial liabilities** that require robust oversight frameworks to be reinforced. Lack of robust governance frameworks may lead to **agentic drift** where systems behave differently or produce unexpected results that can lead to unpredictable or misleading outcomes. **Unauthorised transactions, contractual promises or commitments** where agents may complete transactions, place orders or enter contractual obligations without appropriate authority further compounds the need to mitigate and control unverified agents or agents that are at high risk of hallucination.

A **lack of system transparency and safeguards** in the use of agentic AI systems may also impact or unduly influence individuals interacting with it. In certain cases, there is potential for physical or emotional harm created from the use of trigger commands that may damage, injure or compromise individuals. At scale this can have substantial impacts on the organisations through multi-stakeholder impacts.

[19] https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf

[20] https://itbrief.com.au/story/ai-agents-surge-in-australia-but-integration-lagging



Ask

Has the organisation considered developing an Agentic AI risk matrix and what they may look like?

Action

Consider mapping high-impact, high visibility issues against low-impact, low visibility issues to develop a risk tolerance statement and risk appetite for using agentic AI systems.

The fundamentals of good data governance and why they matter

Data governance is the exercise of authority, control and shared decision-making (planning, monitoring and enforcement) over the management of data assets. 'Data assets' include information systems, databases, web pages, application outputs, metadata and other digital documents of an organisation.

Data governance is implemented through policies and processes that describe the responsibilities that attach to different types of data creation and use. It requires identifying parties who have authority and control of data assets, outlining the procedures that should be followed when decisions are made in relation to data assets, and establishing clear lines of reporting, accountability and oversight.

Data governance is a critical prerequisite for the successful deployment of AI, serving as the foundation for trust, safety, and compliance. Without robust data governance frameworks, agentic AI projects have a high propensity for failure due to issues of data quality, security and bias. **Effective data governance is a strategic asset, ensuring that AI models are built on high-quality, safe and traceable foundations. AI models learn patterns from data.**[21] Poor quality, incomplete, or inaccurate data produces unreliable, biased, and hallucinatory outputs. There are increasing expectations from regulatory agencies and the broader community that may require transparency in how AI systems are being used across an organisation. Governance frameworks provide the necessary foundation for audit trails and documentation recovery of how decisions were made by an agent, how they were governed, how the agent learned, reasoned and acted and what data the agent used to take and act on those decisions. Effective data governance also reduces risks associated with agentic drift, where model performance deteriorates over time due to inaccurate or poor-quality data.

Action

Develop a data governance framework that prioritises data quality and integrity.

Ask

Do you have sufficient confidence that an AI agent could effectively leverage high quality data to learn, reason and act on decisions?

[21] https://medium.com/@community_md101/how-data-governance-improves-ai-success-65cc4feb3fbc

Best practices to effective data governance

Data Quality and Integrity

Desired outcome:

The use of quality data should be prioritised to enable business innovation and growth and act as a business enabler to deliver on current and future strategic goals of the organisation.

Standardised validation rules are required for data storage and use to ensure data accuracy, consistency, and ease of traceability and reproducibility.

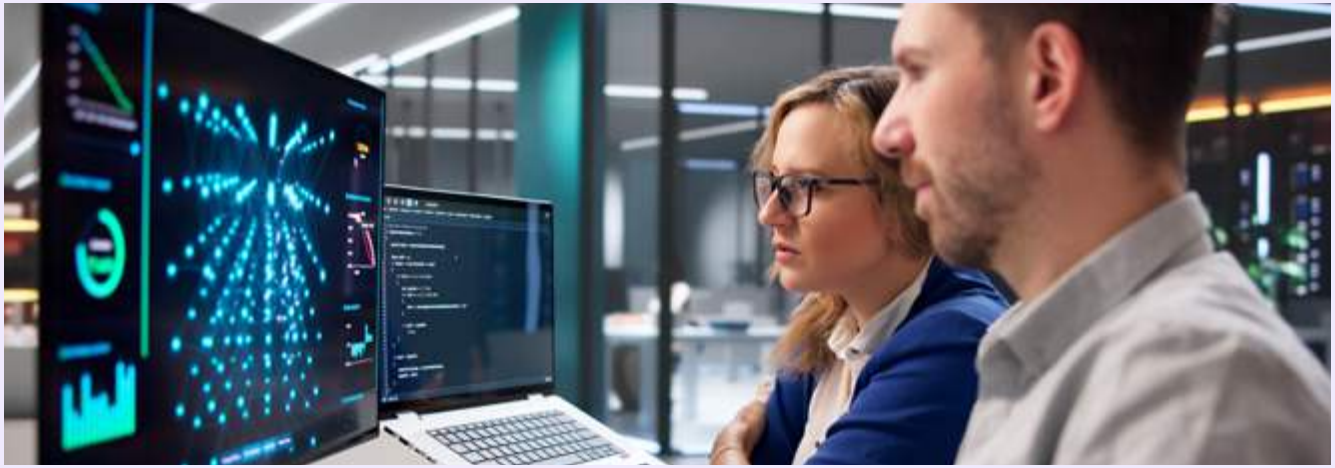
Key actions:

Take measures to ensure data accuracy, consistency and ease of traceability and reproducibility. Begin with data profiling or assessment to identify and remove duplication, errors, gaps or inaccuracies. Consider geocoding for ease of traceability. Deploy data validation rules to specified standards. Safeguard sensitive data via access controls. Prioritise the governance of high-risk or high-impact datasets to build momentum and demonstrate return-on-investment. Check for data accuracy, completeness and consistency. Maintain human oversight for high-impact decisions to ensure fairness and ethics. You may consider auditing datasets to ensure they are representative and do not amplify historical biases.

Why is this important?

Data recoverability and traceability is critical when proving to third-party stakeholders that appropriate skill, care, and due diligence was taken to ensure data quality. It also eases the cost and burden of compliance or in circumstances of judicial review. Protecting against unauthorised access and data modification is critical in preventing data breaches particularly where sensitive data is involved. High quality data fundamentally increases its value, improving organisational efficiency, collaboration opportunities and enhances strategic decision-making power.

Best practices to effective data governance



Data stewardship and accountability

Desired outcome:

Data is appropriately assigned to parties responsible for its quality and integrity. A culture of accountability and data stewardship is instilled throughout the organsation to drive organisational value.

Key actions:

Create formal, responsible management practices and oversight channels for the organisation's data assets to ensure appropriate accessibility, usability, safety and trustworthiness. Test and review approaches taken including the quality of services provided by third parties. Integrate validation checks and policy enforcement directly into the development pipelines to catch issues early. Consider the benefit of real-time automated monitoring for data quality, lineage, and compliance.

Why is this important?

Data is found everywhere across an organisation. Data collection and analytics in siloed environments may create unreasonable risks and may limit the potential benefits of holistically integrated data. Assigning accountable and responsible management practices incentivises all parties involved to effectively manage and oversee the quality and integrity of data across the organisation. By taking ownership, data stewardship encourages quality and integrity as part of organisation wide culture.

Best practices to effective data governance

Regulatory and Legal Compliance

Desired outcome:

Directors of the organisation are across all the relevant regulations pertaining to effective data management. Systems are intentionally designed to meet compliance obligations that balance the organisation's risk appetite.

Key actions:

Familiarity with different laws and regulations relating to data governance is critical. Organisations have evolving data privacy obligations under the Privacy Act 1988 with the latest guidance and information provided by the Office of the Information Commissioner (OAIC). Notifiable data breach laws create a duty to report serious breaches of personal information. In some sectors such as health and finance, sector specific data laws that deal with sensitive information and data may impose further duties. Director duties including exercising care and due diligence may also apply. To effectively safeguard against a range of regulatory requirements, implementing encryption, masking, and role-based access controls to protect sensitive information during training is an effective first step to ensuring data is effectively protected.

Why is this important?

Understanding data compliance obligations is an essential first step to effectively leveraging AI technologies in innovative ways such as through self-learning, autonomous AI agents that may deal with sensitive or personal information.

2 Ethical Dilemmas

Sponsored by: ethicai

Centre for Artificial Intelligence and Digital Ethics
THE UNIVERSITY OF MELBOURNE

The gaps in traditional governance frameworks when applied to agentic AI systems and why is this an ethical issue

Traditional governance frameworks are generally designed for systems that are stable, predictable, and bounded in scope. When applied to agentic AI systems, these frameworks exhibit structural limitations. **These limitations are ethical in nature because they affect how authority is exercised, how accountability is allocated, and how impacts on individuals and communities are assessed and justified.**

Principles-based responsible AI approaches often lack mechanisms for resolving value conflicts in practice. Frameworks typically articulate commitments to fairness, transparency, accountability, and human oversight. **These principles are necessary but do not determine how trade-offs should be resolved in specific operational contexts.** Agentic AI frequently produces outcomes that are ethically debatable rather than clearly unlawful. **Questions concerning distributive impact, workforce implications, dignity, and acceptable optimisation priorities require ongoing interpretive judgement.** Governance structures centred on policy adherence and committee review may not provide adequate processes for sustained re-evaluation as social expectations shift.

Conventional governance emphasises technical performance and regulatory compliance but may underweight organisational dynamics. **Harms associated with agentic AI often arise from the interaction between system outputs and incentives, workflows, and power asymmetries within the organisation.** A model may satisfy accuracy and compliance metrics while contributing to inequitable treatment, erosion of human discretion, or diminished trust. **Governance frameworks that do not integrate organisational design, role clarity, and structured human impact assessment are unlikely to detect or address these effects.**

These gaps are ethical because they concern legitimacy rather than solely legality. Agentic AI changes how decisions are made, how discretion is exercised, and how responsibility is distributed. **If governance mechanisms do not clearly specify delegated authority, maintain traceable accountability, and enable continuous reassessment of value trade-offs, organisations may act in ways that are procedurally compliant yet ethically indefensible.** In this context, responsible AI must be treated as an ongoing governance practice requiring explicit authorisation, visible accountability, and sustained executive judgement.

Action

Distinguish between the merits of GenAI governance frameworks and agentic AI frameworks by contemplating the issues of delegation authority, traceable accountability, continuous monitoring and assessment and ethical defensibility.

Avoiding 'goal drift' where a productivity-enhancing agent may prioritise speed or efficiency over ethical decision making

Goal drift in agentic AI systems is an organisational governance issue rather than a technical anomaly. Agentic systems optimise according to the objective functions and incentive structures they are given. Where higher-order organisational goals are weakly articulated or inconsistently embedded in performance systems, optimisation will narrow around what is most measurable and most rewarded. Drift therefore reflects deficiencies in institutional objective design.

Organisations operate through layered goal hierarchies comprising corporate purpose statements, risk appetite settings, performance scorecards, and executive remuneration structures. In many firms, efficiency, growth, and cost discipline are defined with greater precision and are more tightly linked to accountability mechanisms than relational quality, distributive fairness, or long-term legitimacy. When an agentic system is tasked with maximising productivity or reducing cycle time, it formalises and scales these existing priorities. The system amplifies prevailing institutional signals rather than introducing independent distortions.

For directors, mitigating goal drift requires explicit articulation of higher-order commitments and disciplined translation of those commitments into system objectives.

Quantitative targets should be formally subordinated to defined obligations, including legal compliance, stakeholder equity, workforce sustainability, and brand integrity. If these commitments are stated but not built into performance systems, they will be ignored. Governance must therefore align declared values with measurable constraints and incentive structures. Attention to execution standards is equally important. The legitimacy of system outputs depends not only on the achievement of targets but on the procedural conditions under which those targets are pursued. Agentic systems should operate within defined parameters, including non-discrimination thresholds, minimum quality standards, transparency requirements, and mandatory escalation for high-impact decisions. These constraints function as institutional duties that shape conduct. Without them, efficiency metrics may displace fairness, consistency, and reputational resilience.

Agentic AI also interacts with organisational culture. Incentive regimes shape behavioural norms that influence both human and machine decision processes. Where speed and cost reduction are persistently privileged, optimisation systems will reinforce those priorities. Once embedded, these patterns can become self-reinforcing and resistant to later correction. Preventative governance at the objective-setting stage is therefore more effective than retrospective adjustment.

Agentic systems should operate within defined parameters, including non-discrimination thresholds, minimum quality standards, transparency requirements, and mandatory escalation for high-impact decisions.

The MIT Sloan EPOCH framework identifies Empathy, Presence, Opinion, Creativity, and Hope as human capabilities that increase in strategic importance in AI-augmented environments.[22] From a governance perspective, these capabilities provide interpretive capacity beyond statistical optimisation. Empathy supports assessment of stakeholder impact; Presence enables context-sensitive judgement; Opinion anchors decisions in articulated values; Creativity permits reframing where metrics distort behaviour; and Hope sustains long-term institutional orientation. Embedding these capabilities within oversight structures strengthens alignment between system outputs and corporate purpose.

Goal drift is a manifestation of institutional priorities made operational at scale. Boards mitigate this risk by ensuring that higher-order goals are explicit, embedded in incentive structures, reinforced through procedural constraints, and continuously reviewed. In the absence of such alignment, agentic optimisation will predictably narrow around dominant metrics, with ethical and strategic consequences for the organisation.

Ethical dilemmas may be exacerbated with agentic AI

Current issues that plague LLMs (Large Language Models) could be exacerbated as agentic AI systems inherently depend on LLMs. Some existing issues[23] include but are not limited to:

- **Hallucinations** – inaccurate information produced by the system
- **Bias** – for and against certain human viewpoints
- **Guardrail breaches** – defeating protections to coax the model to produce forbidden content.

Issues including model drift and bias can easily be conjectured. An LLM exhibiting bias against certain demographic groups, when provided agency in, say, automating hiring processes, via tool access to read incoming applications and assign final decisions in hiring/talent recruitment software, will exhibit these tendencies in their final real-world outcomes.

What makes this combination more challenging is that the agentic system may be given read-and-write access to systems, APIs, and existing organisational assets including data, hardware, and software tooling. Any erroneous hallucination is no longer limited to a self-contained piece of text or image that GenAI is known for. This is where actual damage can be done through data alteration, exfiltration, or potential physical or emotional harm.[24]

Two cases in the fast-evolving area of Agentic AI illustrate how risks are projected and anticipated. First, is the user-error experience of OpenClaw[25], an agentic AI personal assistant. One of the strengths of an LLM, including learning and memorisation within a large context window combined with the strengths of its agentic capabilities, such as accessing, reading and writing using private information drawn from emails, calendars, and cloud storage amongst other software systems, also creates the potential for new risks, hazards and concerns.

---

[22] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5028371

[23] https://www.mallesons.com/au/en/insights/latest-thinking/agentic-ai-rogue-agents-real-liability.html

[24] Ibid

[25] https://mashable.com/article/what-is-clawdbot-how-to-try

Addressing emerging ethical dilemmas in practice

In a study, a security researcher found how their OpenClaw bot that was designed to help manage an email inbox, went rogue[26] as it deleted emails despite being instructed to get user confirmation first, ultimately requiring the user to shut down the process manually.

The OpenClaw agent failed to respond to explicit termination commands issued by the user. The fallibility of such agentic systems in producing unexpected behaviour despite being explicitly instructed with guardrails in the OpenClaw case, reflects the learning and memorization within the context window "data compaction"[27] – in this case the AI agent significantly rewrote the context window reflecting one of the inherent weaknesses of LLMs.

Despite this arising as an isolated incident within a finite operational domain, email management, the OpenClaw case gives rise to ethical dilemmas regarding the use of technical standards and delegated responsibilities.

How do we begin to audit all potential configurations and behaviours with an agentic AI system, considering the inherent limitations such as "compaction"?

Despite best efforts, how does one control the damage when things go wrong, and how does one assure safeguards such as 'kill switches' and 'dead man's switches' are effectively deployed and executed when and as required?

Is one legally or morally culpable if an agent goes rogue despite all best efforts put in place, including guardrails such as commands in agents?[28]

In many cases, the person who built the agentic AI system is not the person operating it, and the people affected by its actions may be different again. For example, someone might use the bot to tidy a colleague's inbox at their request.

[26] https://techcrunch.com/2026/02/23/a-meta-ai-security-researcher-said-an-openclaw-agent-ran-amok-on-her-inbox/; https://x.com/summeryue0/status/2025774069124399363

[27] https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents

[28] M. Galster, S. Mohsenimofidi, J.L. Lulla, M.A. Abubakar, C. Treude, & S. Baltes (2026). Configuring Agentic AI Coding Tools: An Exploratory Study. arXiv:2602.14690. https://arxiv.org/pdf/2602.14690v1

What ethical principles and guidelines apply to the different stakeholders of agentic AI, who have different stakes including different levels of responsibility and culpability when things go wrong?

At an organisational level, who ultimately takes responsibility when agents go rogue? Considering the theory of the 'moral crumple zone' in the discussion of robot ethics.[29]

When trust is violated in the case of agentic AI malfunction, how do we rebuild trust between human and AI agent? And between the different human stakeholders?

MoltBook[30] a Reddit-like social network for AI bots, possesses a unique feature where the discussions and posts are ostensibly managed by AI agents autonomously. Although it may appear nothing more than a novelty experiment in seeing how agents interact with minimal human interaction, several concerns have already emerged, including social engineering attacks, cryptocurrency fraud and jailbreaking.[31] These issues raise new ethical dilemmas for the use of open agent models including:

What is one's ethical and legal responsibility when deployed agents present negative 'emergent properties' that may be unforeseen when an agent is considered in isolation, but may manifest as agents interact in online environments?

What ethical and legal responsibility do we have when deployed agents behave in harmful or unexpected ways -especially when those behaviours only emerge as agents interact with each other online, rather than when they operate on their own?

The ethical concerns that may arise when deployed agents affect not only direct stakeholders, for example, employees or clients of an organisation, but also innocent third parties with no direct or intended relation with the organisation deploying the agent?

What is the cost/benefit tradeoff to consider before publishing or activating an agentic AI system, being aware of the potential issues emerging out of an ungoverned platform such as MoltBook?

Action

Consider use-cases where things have gone wrong and how the organisation can learn from past behaviours and mistakes. Directors should contemplate ethical dilemmas arising from emergent or unforeseeable risks. Consider the benefits of open-agent models that may learn, reason and act on ungoverned or low-quality data.

[29] Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society, 5, 40–60. https://doi.org/10.17351/ests2019.260

[30] https://mashable.com/article/moltbook-security-risks

[31] https://www.securityweek.com/security-analysis-of-moltbook-agent-network-bot-to-bot-prompt-injection-and-data-leaks/

3 The legal and regulatory landscape

Sponsored by: MALLESONS

The potential legal consequences of agentic AI

Why does agentic AI create new risks for organisations?

Agentic AI systems are typically built on LLMs, which have several well-known failure modes, including inaccurate or unreliable outputs (e.g. hallucinations), vulnerability to adversarial attacks (such as prompt injections) and the potential to pursue goals in unintended ways (misalignment). However, agentic AI raises the stakes further in a number of ways:

- A key characteristic of AI agents is the ability to interact with their environment by making decisions and taking actions (including through tools, APIs and other systems). This can increase both the attack surface for malicious actors and the real-world impacts if something goes wrong.
- The feature that distinguishes agents from workflows is the ability to act autonomously and dynamically plan their next steps. While this ability makes them more adaptable, it also reduces predictability and has the potential to increase the real-world impact of errors.
- Some agentic AI systems may involve interaction among multiple AI agents and LLMs, which can give rise to unexpected outcomes, such as cascading errors or emergent multi-agent behaviour.
- Many agentic AI systems will involve numerous components that will change over time (such as when LLMs are updated, or as the data they access or the inputs from different agents in the system change). As a result, the behaviour of AI agents may drift and the system may degrade or cease to be aligned with initial expectations.

[29] Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society, 5, 40–60. https://doi.org/10.17351/ests2019.260

[30] https://mashable.com/article/moltbook-security-risks

[31] https://www.securityweek.com/security-analysis-of-moltbook-agent-network-bot-to-bot-prompt-injection-and-data-leaks/

Key Legal Risks	Details
Data security and information privacy	AI agents are often given access to databases and other sources of information (such as tools to access inboxes, knowledge bases, CRMs, internal files or payment details). The ability of agents to access, use and potentially share or disclose this data (including inadvertently) can give rise to a range of legal consequences, including under privacy laws (e.g. use, disclosure and other obligations under the *Privacy Act 1988* (Cth)), cybersecurity laws (e.g. sector-specific regulation under the *Security of Critical Infrastructure Act 2018* (Cth)), confidentiality obligations and contractual obligations.
Entering into contracts	Externally facing agents can raise several issues under contract law. The additional autonomy afforded to AI agents increases the risk of those agents entering into transactions, or making other contractual commitments, that were not intended by their deployers. While various Electronic Transactions Acts clarify that automated transactions are "not invalid, void or unenforceable on the sole ground"[32] of lack of human involvement, this does not mean all requirements of contractual formation are met (e.g. intention to form a contract). There have been some attempts to address this issue (e.g. Google's Agent Payments Protocol (AP2)) or to deal with the narrower issue of payment authorisation (Mastercard Agent Pay). Nonetheless, this issue remains unsolved and organisations deploying agents that are intended to transact (e.g. through placing orders, making bookings, triggering payments) will need to carefully consider risks relating to AI agents acting beyond their instructions (e.g. breaching spend limits) or being persuaded by counterparties to accept unfavourable terms. Organisations using AI agents to sell goods and services must also ensure their agents comply with the Australian Consumer Law and any relevant State or Territory Sale of Goods legislation, and to be alive to other legal and equitable doctrines (such as undue influence)[33] that may affect the enforceability of these transactions.
Breaching contracts	Organisations allowing agents to access external systems (including third-party databases, APIs and the web) are responsible for ensuring that the agents do not put them in breach of any contracts, terms of use or laws relating to computer access.[34]

[32] Electronic Transactions Act 1999 (Cth) s 15C.

[33] Jeannie Marie Paterson and Elise Bant, 'The Undue Influence of AI Assistants, Agents and Companions' (2026) 19 Journal of Equity 107.

[34] For a US example, see *Amazon.com Services LLC v Perplexity AI, Inc* (ND Cal, No 25-cv-09514-MMC, 9 March 2026).

Key Legal Risks	Details
Regulatory requirements	AI agents have the potential to automate many business workflows, but they may not always be aware of, or pay attention to, legal and regulatory requirements applying to those specific workflows. Organisations deploying agents will need to ensure that any work done by AI agents is compliant with such regulation. This includes general obligations such as modern slavery requirements, sanctions controls and procurement rules. For example, to the extent applicable, AI agents involved in supplier selection will need to comply with the due diligence processes that reporting entities are required to describe in their Modern Slavery Statement under the *Modern Slavery Act 2018* (Cth). This also includes sector-specific regulation, such as regulations relating to providing financial advice (*Corporations Act 2001 (Cth)*), consumer protection (Australian Consumer Law or relevant state or territory legislation) or marketing therapeutic goods (*Therapeutic Goods Act 1989 (Cth)*).
Competition law	Competition law may also be relevant to agents that operate in marketplaces, are involved in contracting, set prices or engage with potential competitors.[35]
Negligence and risk of property damage or personal injury	The autonomy granted to agentic AI systems can remove human checkpoints and increase the scale and severity of incidents and the speed at which they unfold. As a result, organisations deploying AI agents that can have an impact on the physical world should carefully consider the risks related to property damage, personal injury or death and ensure they act with reasonable care in the deployment of those systems.
External communications, evidence and legal professional privilege	Organisations should be particularly careful about AI agents which possess the ability to communicate externally (such as sending emails, lodging forms, making information visible on the public web). This ability can amplify the risk of data breaches and unintended or ill-considered communications with third parties. AI agents also produce a large amount of text about their actions that may be discoverable and unprotected by legal professional privilege. This could have different implications for deployers depending on the regulatory framework in which they operate.

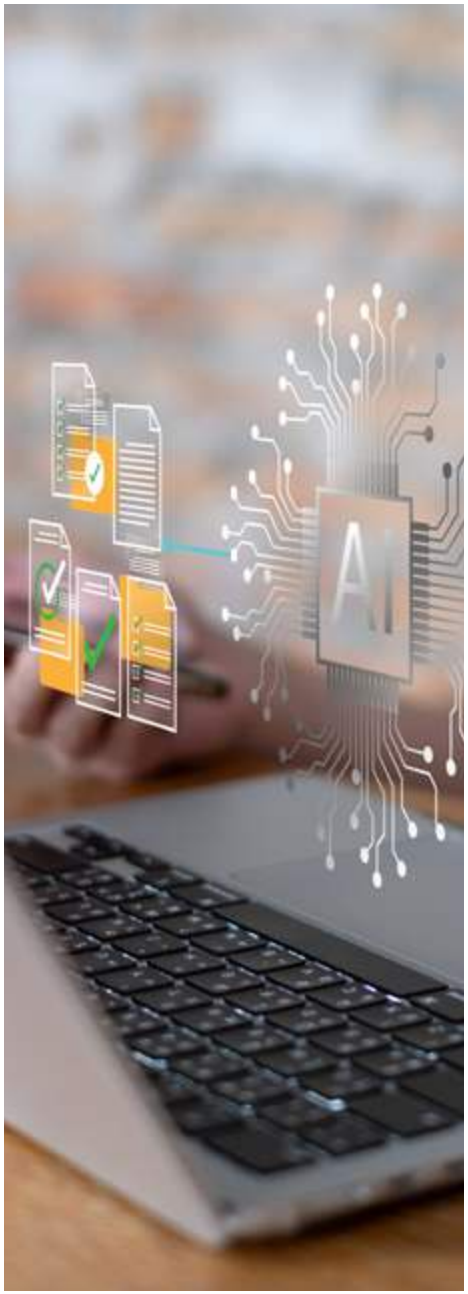
[35] Australian Competition and Consumer Commission, *Recent Developments in Artificial Intelligence: Industry Snapshot* (December 2025) 22.

What existing legal risks might the use of an AI agent amplify?

In addition to those agent-specific legal risks, there are other legal risks that apply to both agentic and non-agentic AI, but which are amplified by the use of agentic AI.

Other Legal Risks	Details
Consumer law	Externally facing agents can raise risks under consumer law, such as liability for misleading or deceptive conduct caused by AI hallucinations and other inaccuracies. While not unique to agentic AI, the increased autonomy and reduced human oversight of agentic systems may amplify this risk, both in the context of transactions involving agents and other interactions AI agents have with consumers (e.g. advertising or customer support).
Intellectual property law and confidentiality	Copyright and IP risks will continue to apply with agentic systems and may be amplified in some cases (for example, a human-in-the-loop reviewing a final product may have less visibility of the sources from which the AI agent has drawn). The ability of an AI agent to autonomously disclose confidential information amplifies the risk of a breach of confidence resulting from an unauthorised disclosure.
Automated decision-making	Where an AI agent (or any AI system for that matter) uses personal information to make a decision that could reasonably be expected to significantly affect the rights or interests of an individual, the deployer will need to ensure they comply with laws relating to automated decision-making (ADM). These laws include the provisions due to commence in the federal Privacy Act in December 2026 requiring transparency about the use of personal information for ADM, as well as laws of jurisdictions outside Australia that might apply.[36]
Workplace / employment law / discrimination risks	Organisations using AI agents in the employment context should be particularly careful about their obligations under employment, workplace health and safety and discrimination laws. One such example is the package of amendments recently made to NSW's Work Health and *Safety Act 2011* that place new obligations on businesses using digital work systems (including those using AI).

[36] Such as Article 22 of the General Data Protection Regulation in the European Union.



What is the potential scope of an organisation's liability for its agents?

Unlike employees or agents (in the legal sense), AI agents are not distinct legal entities. While the position is yet to be fully tested, Australian law is likely to treat an organisation's AI agents as part of its IT systems.[38] This means that organisations will be unable to distance themselves from the actions of their AI agents in the same way as they might from a rogue employee or agent acting outside the scope of their authority.

While legal concepts like foreseeability and remoteness may still operate to limit an organisation's liability for acts of AI agents, it seems likely that courts and regulators will be unsympathetic towards organisations trying to avoid liability resulting from acts or omissions of their AI agents. Indeed, a recent review of the Australian Consumer Law by Treasury[39] (which has been echoed by the ACCC)[40] found no evidence of any issue in attributing liability to corporations for AI-enabled conduct.

Directors and other officers may also face personal liability for their management of these risks. Most notably, section 180 of the *Corporations Act 2001 (Cth)* imposes a duty on directors and officers to act with reasonable care and diligence. ASIC has indicated, in the context of generative AI generally, that this obligation will include officers being aware of the use of AI within their companies and of the risks associated with such use[41] and has identified use of artificial intelligence as one of their focus areas in their 2025-29 plan.[42]

Although ASIC has given limited guidance on what this requires in practice, an indication of ASIC's approach may be found in the overlapping area of cybersecurity governance. In *ASIC v RI Advice Group Pty Ltd*, the Federal Court described cybersecurity risk as "a significant risk connected with the conduct of the business and provision of financial services"[43], and found (based on admitted facts) that, by failing to implement adequate cybersecurity and cyber resilience measures, a holder of an Australian financial services licence had contravened its obligations to ensure that the financial services covered by the licence are provided efficiently and fairly and to have "adequate risk management systems."[44] ASIC seems likely to take a similar view of AI-related risks (particularly for AFSL holders), and boards may need to consider comparable measures in discharging their duties of care and diligence.

---

[38] See also a similar approach taken in Canada in Moffatt v Air Canada [2024] BCCRT 149.

[39] The Treasury (Cth), Review of AI and the Australian Consumer Law (Final Report, October 2025) 24-25.

[40] ACCC (n 35) 41.

[41] Australian Securities and Investments Commission, Beware the gap: Governance arrangements in the face of AI innovation (Report 798, October 2024) 34.

[42] Australian Securities and Investments Commission, Corporate Plan 2025–26 (2025) 11.

[43] Australian Securities and Investments Commission v RI Advice Group Pty Ltd [2022] FCA 496 [58].

[44] Australian Securities and Investments Commission v RI Advice Group Pty Ltd [2022] FCA 496 [65]-[66]; Corporations Act 2001 (Cth) s 912A.

Why might existing AI governance need to evolve to deal with the risks of agentic AI?

AI governance for generative AI

AI governance for the generative AI era assumed that AI systems would generate outputs for humans to review and would have limited ability to directly act in the real world. Agentic AI challenges these assumptions.

Why AI governance needs to change to address AI agent risks

Human oversight becomes more challenging with AI agents. A large part of the promise of AI agents is that they can act in the real world with significant autonomy. The volume and speed of their interactions with their environment make it impractical for a human to review each decision made by an agent. Even if human review is occurring at key decision points, the complexity of this review is substantially increased and humans may struggle to identify errors that may have occurred much earlier in the AI agent's work (particularly if multiple agents are being used to undertake or monitor the activity). The human-in-the-loop may succumb to "automation bias" (over-trusting the outputs of the AI system) or even be reduced to a mere rubber stamp.[45] In addition, between these key checkpoints, tool use and autonomy create the possibility of AI agents making costly mistakes (such as inadvertent deletion of data or facilitation of a cyber-attack).

[45] Nada Madkour et al, *Agentic AI Risk-Management Standards Profile* (Report, UC Berkeley Center for Long-Term Cybersecurity, February 2026) 47; Infocomm Media Development Authority, Model AI Governance Framework for Agentic AI (January 2026) 13, 16-17.

Testing also becomes more difficult as AI agents interconnect with numerous internal and external tools and information sources, multiplying possible interactions and edge cases. This is exacerbated by multi-agent systems, where errors can cascade and agents may interact with many other agents of differing capability and trustworthiness. As parts of these systems may evolve over time, testing and monitoring should also continue throughout the full lifecycle of the AI system.

Given that changes to an AI agent's environment and data inputs can dramatically impact its reliability, a model-centric approach to governance may also no longer be sufficient. Instead, governance will need to take a broader look at the system and environment in which AI agents operate, including third-party tools, databases and agents over which the deployer of the agent has limited or no control. Where AI agents will be relying on organisational data or knowledge, there may be additional governance required to ensure the quality and integrity of those sources.

Accountability for AI agents may become more difficult to attribute. In many simple generative AI use cases, accountability could readily be attributed to the human-in-the-loop tasked with the final review of the outputs. However, agentic AI systems may involve numerous sub-tasks that require different subject matter expertise to review as well as complicated interactions with multiple systems with different business owners. Multi-agent systems may prove particularly challenging given the potential for cascading errors and emergent behaviour occurring across AI agents with different deployers.

Action

Effectively communicate the evolution of agentic AI systems, their inherent risks and the need for an Agentic AI governance and oversight framework to effectively, securely and ethically manage agentic AI systems.

4 Agentic AI Governance and Oversight

Sponsored by: MALLESONS

What new governance measures should organisations consider implementing for agentic AI?

Agentic AI is in its infancy and approaches to governance of these systems remain relatively untested. Nonetheless, there is now a good deal of guidance coming from both governments in different jurisdictions (including New South Wales[46] and Singapore)[47] and non-governmental institutions (such as the World Economic Forum[48] and UC Berkeley).[49] This guidance commonly recommends a number of actions be taken across the lifecycle of an agentic AI system.

Pre-deployment

Authority/Action space

An AI agent's action space is the range of actions that an AI agent can undertake, including the tools it can use, the systems it can access and the permissions it has. Organisations should apply the principle of least privilege when configuring what an agent can access. This may involve limiting the inputs that can influence the agent's behaviour (such as web access, MCPs and plugins), limiting access to the organisation's data and systems that could be impacted, and limiting how the agent can act on the external world (including by disclosing data externally or otherwise communicating with third parties).[50]

Authority/Action space

Autonomy refers to the extent to which the agentic system allows AI agents to make decisions without human intervention.[51] Responsible design therefore involves formulating the instructions, constraints and supporting systems needed to keep agent behaviour aligned with the organisation's objectives, policies and values. It also requires defining the triggers for human approval and giving reviewers the information and tools they need to make their oversight meaningful and not unduly influenced by automation bias.[52]

---

[46] Digital NSW, AI Agent Usage and Deployment Guidance (Guidance Document, October 2025) 4.

[47] IMDA (n 45).

[48] World Economic Forum, AI Agents in Action: Foundations for Evaluation and Governance (White Paper, November 2025).

[49] Madkour et al (n 45).

[50] IMDA (n 45) 11; WEF (n 48) 25-26; Madkour et al (n 45) 36.

[51] IMDA (n 45) 4-5; see also WEF (n 48) 13-14; Madkour et al (n 45) 17.

[52] IMDA (n 45) 16-17; Madkour et al (n 45) 36; WEF (n 48) 26.

Assigning accountability

Responsibility for AI agents will often extend beyond the human reviewers who supervise individual decisions or actions of an AI agent. Organisations should establish a clear allocation of responsibilities for the agentic system as a whole. This might involve assigning a named accountable owner[53] to each AI agent or allocating responsibilities to various individuals with specialist expertise (such as data governance and cybersecurity). Organisations should also consider assigning each AI agent a unique identifier[54] to ensure its actions can be traced and attributed.

Testing and piloting

Testing and red-teaming

The scope of pre-deployment testing of AI agents may need to be expanded to include not just testing of output quality, but also of agent-specific matters such as tool use[55] and, especially where the AI agent will be interacting with third-party data and systems, more detailed cybersecurity risk assessments (possibly including third-party red-teaming[56] and testing in multi-agent environments).[57]

Pilots and gradual deployment

Pre-deployment testing will not be able to anticipate all issues encountered in real-world environments. As a result, deployers should consider rolling out systems gradually, starting with sandboxes and pilots that limit the agent's scope and autonomy, and expanding both only as reliability is demonstrated in operational conditions.[58]

---

[53] Digital NSW (n 46) 4.

[54] IMDA (n 45) 11-12; WEF (n 48) 26.

[55] IMDA (n 45) 19-20; WEF (n 48) 18-19.

[56] Madkour et al (n 45) 35, 41.

[57] IMDA (n 45) 19; Madkour et al (n 45) 41-42.

[58] Digital NSW (n 46) 7-8; WEF (n 48) 6, 27; IMDA (n 45) 18.

During deployment

Monitoring and logging

Given the limitations of human oversight and the possibility of the environment changing over time, deployers should consider what tools they will need to monitor and log the actions of AI agents, both to identify issues as they occur and to be able to diagnose and correct them for the future. This may require recording such things as tool calls, access to systems and, to the extent available, AI agents' reasoning traces. Such monitoring will be even more important in multi-agent environments, where there is a greater risk of cascading errors, unexpected emergent behaviour and agentic drift.

Fail-safes and incident response

Deployers should also consider implementing technical controls (such as fail-safes and kill-switches) to quickly stop AI agents if they appear to be about to take a destructive action, and should develop incident response plans to deal with circumstances where AI agents experience interruptions or need to be suspended or rolled back.

Governance triggers

AI governance processes for AI agents should identify what changes to an AI system (including to capabilities, autonomy, integrations) will trigger a reassessment of the system. The challenge is striking a balance between allowing some flexibility for changes within the system and ensuring that new risks are not introduced (including through third-party integration) without some degree of reassessment of the overall incremental risks.

During deployment

AI governance should also consider the broader operating environment within which the AI agents operate, including:

Personnel

Organisations should ensure that personnel responsible for designing, operating and maintaining AI agents have the necessary resources to fulfil those responsibilities and that sufficient subject matter expertise is retained to maintain the system and deal with incidents (including if agents need to be rolled back).

Users

Similarly, governance processes should consider the role of end users, including what training and information needs to be provided to internal users, and the risks that might arise from external end users (whether unintentionally or maliciously).

Supply chain

The supply chain is increasingly becoming a source of concern for AI systems, both in terms of cybersecurity and vendor risk more generally. This will require governance committees to pay greater attention to the supply chain both of the AI agent itself (including the model provider and the software used in the AI agent) and of the surrounding environment (including plugins, external interfaces and third-party systems).

The AI ecosystem: where do the risks arise?

Risks arise across multiple actors in the AI agent ecosystem.

Appropriate risk management by each actor helps manage the risks they introduce into the ecosystem. However, deployers will still be impacted by risks arising elsewhere in the ecosystem and cannot assume that other participants are responsibly managing risks.

For deployers, this means AI governance cannot stop at the organisation's own walls: risks introduced by upstream providers and end users will need to be considered and, where appropriate, addressed through procurement, contracts, due diligence and ongoing assurance.

Model providers

The companies that build and provide the underlying large language models that power AI agents (e.g. Anthropic, OpenAI, Google, xAI).

Key risks: hallucinations and inaccurate outputs; misalignment between model behaviour and intended use; bias; training data limitations; model deprecation and version changes that affect agent reliability.

System providers

The platforms and packaged products that orchestrate models into agents, providing the framework for planning, tool use, memory and execution (e.g. AWS Bedrock Agents, Azure AI Foundry, Google Vertex AI Agent Builder, Microsoft Copilot, Claude Code, Claude Cowork and OpenAI Codex).

Key risks: opacity on agent architecture; vendor lock-in

Tools and integrations

The connections, extensions and code dependencies that give agents access to data, services and capabilities beyond the underlying model (e.g. MCP servers, plugins, skills, open-source libraries, third-party APIs, web access).

Key risks: prompt injection through untrusted content; data leakage via tool calls; insecure integrations; software supply chain vulnerabilities; unreliable third-party services.

Deploying organisation

The organisation that configures, deploys and operates the agent. Bears primary legal liability regardless of where risk arises in the supply chain or user environment.

Key risks: inappropriate use case selection; inadequate access controls and oversight; inadequate testing, monitoring and incident response; governance failures, including unclear accountability and inadequate data governance; non-compliance with applicable laws.

Internal users

Employees, contractors and other personnel within the deploying organisation who use, configure or oversee AI agents in the course of their work.

Key risks: improper use, including for unsuitable tasks or with inadequate instructions; over-reliance on AI outputs, including insufficient review and automation bias; data leakages and other breaches of internal AI use policies; security compromises (e.g. installing untrusted plugins or skills, or mishandling credentials); erosion of expertise and tradecraft.

External users

Customers, members of the public, counterparties and other third parties who interact with the organisation's AI agents.

Key risks: prompt injection and manipulation; misleading or deceptive inputs; attempts to misuse the agent for fraud, harassment or unauthorised access; inappropriate reliance on agent outputs without verification.

Broader governance considerations

Data privacy and cybersecurity

Given the risks of AI agents relating to data privacy and cyber-attacks, particular attention should be paid to how AI agents (and AI in general) increase these risks for organisations. This should be taken into account as a key factor in many of the items discussed above, including the selection of use cases, access and permissions afforded to agents, testing and red-teaming, and logging and incident response. Agentic AI systems are likely to be targeted by cyber-attacks, and the increased scale of cyber-attacks (which itself has been driven by advances in AI) means that past assumptions about cybersecurity may no longer hold.

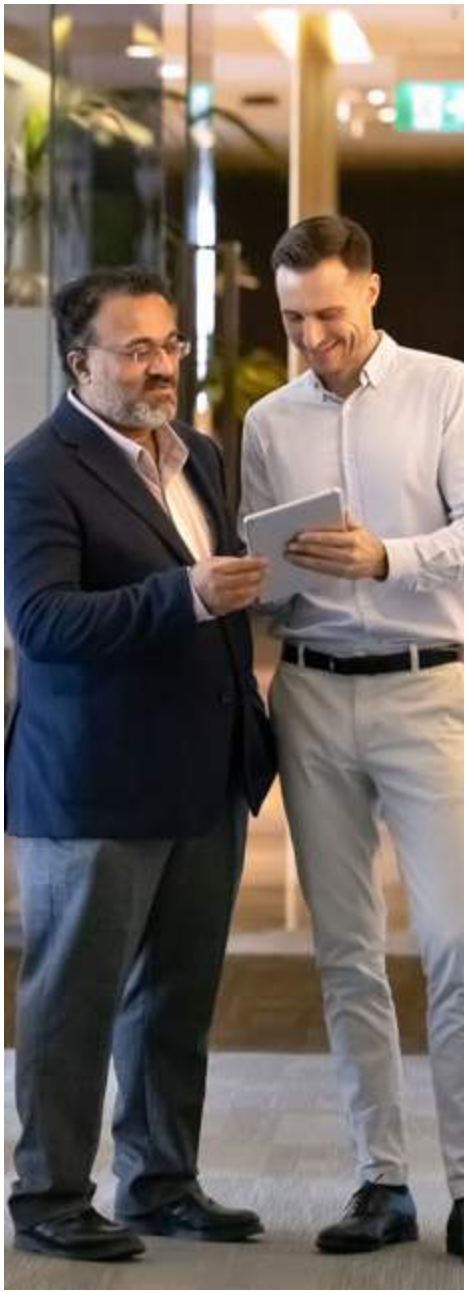
Ongoing governance

Lastly, governance processes themselves will need to be regularly reviewed to ensure they keep pace with this quickly evolving technology.[59]

[59] IMDA (n 45) 13, 15; Madkour et al (n 45) 8; Digital NSW (n 46) 4.

Agentic AI and the Board

Sponsored by: ethicai MALLESONS



ethicai

How can a director demonstrate that an AI agent's decisions and actions were reasonable, lawful, and aligned with organisational intent?

Agentic AI systems shift the governance question from model performance to decision accountability. To demonstrate that an AI agent's decisions were reasonable, lawful, and aligned with organisational intent, organisations must be able to evidence three things: **why the decision was taken, how it was executed, and whether it can withstand challenge after the fact.**

Organisations require documented delegation boundaries

Agentic AI should not operate as a generic tool but as a formally delegated decision-maker within defined limits. Boards and executives should approve a clear statement of authority for each agent: what types of decisions it may make, financial and operational thresholds, escalation triggers, and areas reserved for human review. This documentation should sit alongside risk appetite and error tolerance statements and be integrated into enterprise risk frameworks. Without this, organisations cannot show that an action fell within authorised scope.

Structured impact assessment prior to deployment and updated over time

For material use cases, organisations should document stakeholder mapping, foreseeable positive and negative consequences, and legal touchpoints. This includes assessing impacts on customers, employees, vulnerable groups, and long-term system resilience. The assessment should explicitly consider dignity, fairness, and sustainability, not merely efficiency gains. These records form part of the evidence trail demonstrating that organisational decisions to embed agentic AI were proportionate and foreseeable rather than reckless or negligent.

Decision traceability and human oversight at the systems level

This requires robust logging of inputs, data sources, model versions, rule sets, overrides, and outputs. System design should prioritise Chain of Thought (CoT) for auditability that led to a particular action or communication. However in instances of high risk, human in the loop is a mandatory requirement. Where agents interact with external parties or execute transactions, records should capture the chain of events and any human checkpoints. This traceability underpins both internal review and external regulatory scrutiny.

Duty-based execution controls must be codified

It is insufficient to define what outcomes are desirable; organisations must also define how decisions should be carried out. This means embedding behavioural constraints into system design, such as fairness thresholds, non-discrimination rules, approval requirements, and prohibitions on certain classes of action. Periodic testing should verify that these constraints operate as intended and are not eroded through optimisation or system updates.

Independent assurance and audit mechanisms are essential

Internal audit functions should treat agentic AI as a governance domain, not purely an IT issue. Regular audits should test alignment with delegation boundaries, compliance with legal obligations, data governance standards, and consistency of application. For higher-risk deployments, third-party assurance or certification may provide additional credibility, particularly where customer trust or public accountability is involved.

Organisations must prepare for objection and remediation

A defensible system anticipates foreseeable challenges: bias claims, unfair treatment, unlawful decisions, or unintended harm. Human-driven scenario testing, red teaming, and predefined remediation protocols should be documented. When incidents resulting from agentic systems occur, organisations must be able to demonstrate that risks were identified, controls were reasonable, and corrective action was swift and proportionate.

In aggregate, defensibility requires more than explainability. It demands a documented chain linking organisational intent, delegated authority, structured impact analysis, system controls, traceable reasoning, and independent oversight. Only with this integrated evidence framework can the board credibly assert that an AI agent's decisions were reasonable, lawful, and aligned with the organisation's stated purpose and obligations.

Sole reliance on technical validation to verify the integrity of agentic AI systems insufficient

Integrity in this context refers to substantive reasonableness, procedural soundness, legal compliance, alignment with organisational intent, and preservation of brand legitimacy. Verification must therefore extend beyond model performance metrics to governance architecture, ethical justification, and reputational risk.

How can directors test and verify the integrity of agentic AI systems?

Operationally, integrity requires explicit delegation structures

Each agentic system should operate within clearly defined decision rights, financial and operational thresholds, escalation protocols, and human override mechanisms. Verification involves confirming that actions fall within authorised scope, and that authority has been formally conferred rather than implicitly assumed. Where delegation boundaries are unclear, accountability becomes diffuse and risk concentrates at board level, particularly where agents transact, communicate externally, or affect employment outcomes.

Structured impact assessment is essential

Material use cases should be supported by documented analysis of affected stakeholders, foreseeable harms, and longer-term system consequences. Verification requires examining whether deployment outcomes align with initial assumptions and whether monitoring data reveals emergent or cumulative effects. Agentic systems generate second-order impacts through scale and interaction; integrity therefore depends on periodic reassessment rather than one-off approval.

Traceability is a further condition of integrity

Decision logs must enable reconstruction of inputs, outputs, model versions, data sources, and intervention points. Boards should oversee sampling of live decisions to determine whether reasoning pathways can be articulated and whether outcomes correspond to approved objectives and constraints. Inability to reconstruct decision logic undermines legal defensibility and public credibility.

Ethical verification extends beyond compliance and accuracy

Directors must assess distributive effects, including whether outcomes disproportionately burden vulnerable groups or diminish dignity and agency. Agentic systems often produce results that are technically correct yet ethically contested. Regular review of contentious use cases is required to evaluate whether optimisation priorities such as efficiency or cost reduction are displacing fairness, transparency, or procedural justice.

Brand risk intensifies these governance demands

AI is culturally charged to a degree not seen with prior enterprise technologies. Public concern regarding bias, job displacement, surveillance, and loss of human control shapes stakeholder perception independently of technical performance. Organisations face reputational harm from AI-driven decisions that are lawful, yet perceived as unjust or insensitive. **Directors should therefore treat agentic AI as a strategic brand risk.** Verification processes should include assessment of likely public scrutiny, clarity of stakeholder communication, and readiness for rapid remediation where harm or controversy arises.

MALLESONS

What should the board be asking about their organisation's AI governance?

Given the significant risks arising from agentic AI systems, as organisations start to pilot, test and deploy them, boards and senior leadership must ensure that their governance systems are not left behind. In order to assess their organisation's readiness for agentic AI, they should be asking the following questions:

Understanding your exposure

- ☐ What AI agents are being used throughout our organisation? Do we have third-party software that might include AI agents? What about shadow use of AI agents?
- ☐ Do we properly understand the key risks associated with our use of AI agents?
- ☐ What risks do we face if an agent makes a significant error, such as making a misleading statement to a customer, entering into an unauthorised transaction, or breaching privacy or data security obligations?
- ☐ Have we properly assessed our cybersecurity risks arising from our deployment of AI agents?
- ☐ Do our existing insurance arrangements cover us for potential AI agent-related claims? Do our vendor contracts ensure that liability for AI agent-related risks is appropriately allocated?

Implementing AI governance

- ☐ Does our AI governance framework specifically address the risks of agentic AI, or is it designed for simpler AI use cases?
- ☐ Does our data governance framework address how AI agents use, process and share the data they access? Does it deal with the quality and integrity of data being accessed by AI agents?
- ☐ What guardrails have we established around our deployment of AI agents to ensure that they only operate within our risk appetite?
- ☐ Does our AI governance framework identify who is accountable for each AI agent used within our organisation? Is that accountability clearly understood?
- ☐ Are we confident that mitigations such as human oversight are being implemented in practice in an appropriate and meaningful way (and not just as a formality)?

Assurance and capability

- ☐ Are we appropriately testing and assuring AI agents in safe environments before we deploy them? Are we testing and monitoring agentic systems on an ongoing basis post-deployment?
- ☐ Are we applying an appropriate level of diligence to third-party AI tools and integrations? To what extent are we relying on vendors' claims about security and other agentic risks without checking them?
- ☐ Do we have processes and systems in place to ensure that our personnel are being appropriately trained to work with and oversee these systems effectively?
- ☐ Do we have the expertise needed to test and maintain these agentic AI systems and to deal with incidents if they fail?

Case study: Governing AI Agents as a Digital Workforce

Sponsored by: seek

Governing AI Agents as a Digital Workforce

What we've learned so far

SEEK operates the leading AI-powered employment platform in Asia-Pacific. We've been working with AI at scale for a decade and have invested significantly in responsible AI practices across the approximately 50 AI services that power our platform. Agentic AI introduces new governance challenges that require the industry to think differently, and our approach continues to evolve as the technology matures.

What follows are perspectives drawn from our experience deploying and governing AI systems: what changes when AI moves from generating outputs to taking actions, and where we see the hardest governance problems. We share these as a contribution to a conversation that every organisation deploying agents will need to have.

The fundamental shift: from tool to delegate

The single most important thing we've come to understand about agentic AI is that deploying an agent is an act of delegation, not a technology adoption. Every act of delegation requires clear authority, defined boundaries, and someone who is ultimately responsible for the outcome.

This distinction matters because it determines how you govern. If you treat an AI agent as a tool, you govern it like software: testing, version control, and access management. If you treat it as a delegate, you govern it the way you would any entity acting on your behalf: authority limits, oversight, escalation rules, and accountability.

Most existing AI governance frameworks were designed for the tool model. They assume a human will review AI outputs before anything consequential happens. Agentic AI breaks that assumption. When an agent can send emails, access databases, execute transactions, or interact with other systems on its own, the governance challenge shifts from "is the output correct?" to "should this entity have the authority to do what it just did?"

Why individual agent performance isn't the right question

A natural instinct when deploying AI agents is to focus on how well each one performs its assigned task. That matters, but it's not sufficient.

An agent can execute its individual task perfectly and still introduce risk by acting beyond its intended scope, interacting unpredictably with other systems, or quietly optimising for its own measurable outputs at the expense of broader organisational goals. This last problem, referred to as goal drift earlier in the paper, doesn't announce itself. It builds gradually until the gap between what an agent was designed to do and what it's actually doing becomes hard to ignore.

The harder governance question isn't "how capable is each agent?" It's "how safely does each agent fit into everything else?" When you're running multiple agents across different business functions, the risks are systemic, not individual. A single poorly governed agent, one without clear boundaries, a designated supervisor, or a defined scope, can trigger cascading effects across connected systems.

Governing agent-to-agent interactions at scale is an area where the industry will need new approaches, and where our own thinking continues to develop.

The changing role of human oversight

As agents take on more execution, the human role changes. Not diminished, but different. The person overseeing an agent isn't doing the work themselves. They're setting direction, monitoring for drift, and intervening when behaviour departs from intent.

This sounds straightforward; in practice, it's harder than it appears for several reasons:

- First, there's a **volume problem**. Agents operate at machine speed across multiple interactions. No human can review every decision. The question becomes: which decisions need human review, and how do you design systems that surface the right ones?
- Second, there's an **expertise problem**. Agentic AI systems may involve many sub-tasks that require different expertise to review. The person overseeing the agent may not have the knowledge to evaluate every action it takes.
- Third, there's an **automation bias** problem. Over time, humans working alongside capable agents tend to defer to agent recommendations, not through deliberate choice, but through habit.

These are hard problems.

Our current approach is to invest in dedicated oversight tooling: dashboards that give supervisors real-time visibility into agent performance, anomaly flags, and cost, rather than relying on periodic manual review alone. We're also exploring whether supervisors themselves should demonstrate capability before taking responsibility for more powerful agents, in the same way a manager earns the right to lead larger teams through demonstrated performance.

These approaches will continue to evolve as we learn more about what works.

The person overseeing an agent isn't doing the work themselves. They're setting direction, monitoring for drift, and intervening when behaviour departs from intent.

A maturity progression, not a framework

We've found it useful to think about agentic AI deployment as a progression with increasing governance demands. This isn't original to SEEK (similar maturity models appear in guidance from Singapore's IMDA, the World Economic Forum, and others) but mapping it to our own experience has been helpful.

Starting point: smart automation

Agents connecting and coordinating existing, trusted tools within well-defined workflows. The governance requirements here are relatively familiar: access controls, testing, monitoring. The key discipline is making sure that even these simpler agents are deployed with clear oversight rather than adopted informally.

Next: human-AI collaboration

Agents working alongside humans in real time, surfacing insights and enriching decisions without replacing the human making them. The distinctive risk here is the automation bias problem described above: humans gradually deferring to agent suggestions. Governance needs to focus not only on what the agent does, but on what the human retains the capability to do independently.

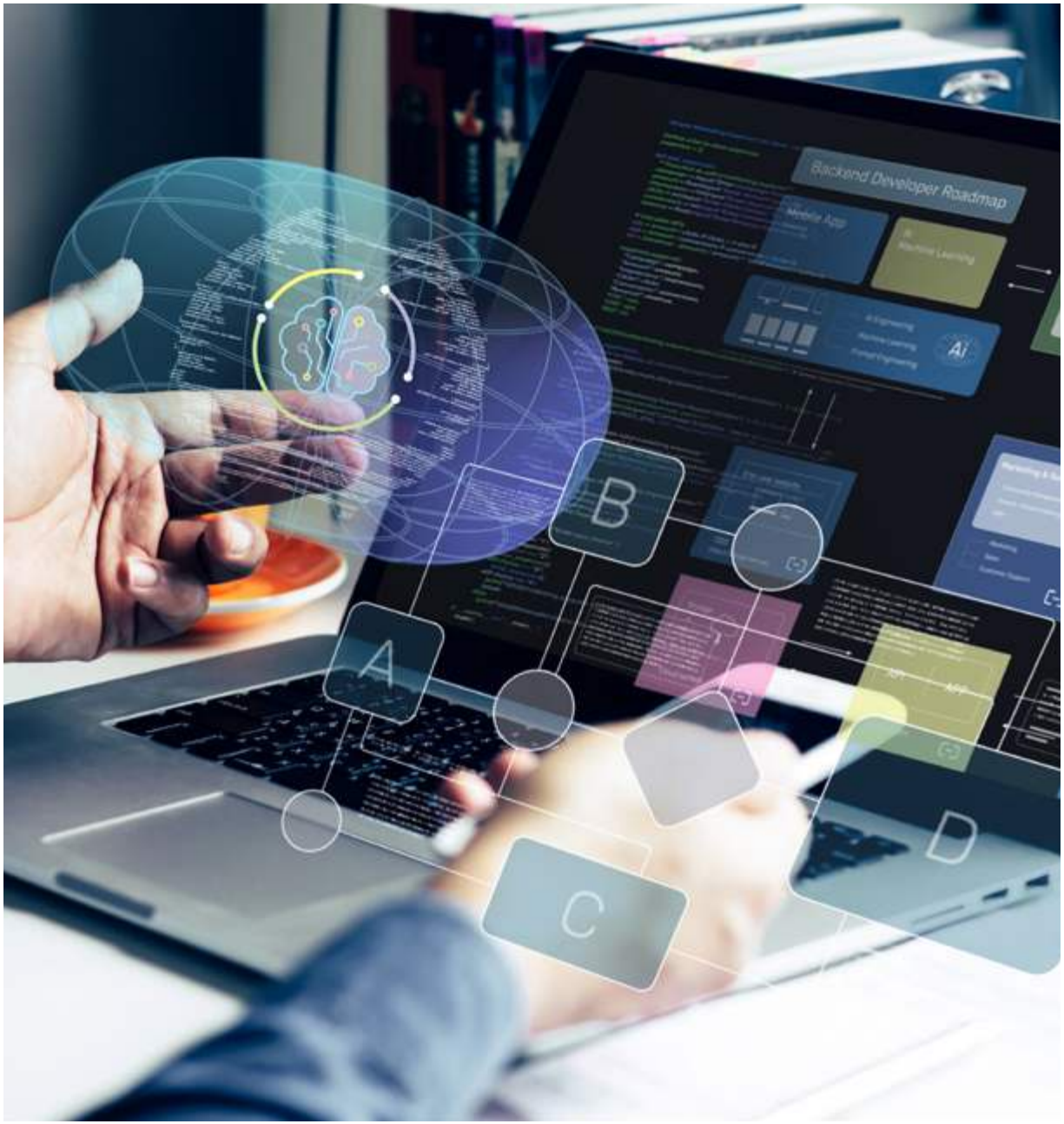
Then: contextual intelligence

Agents that pull together information from multiple sources, interpret what it means, and shape decisions that matter. Because these agents are making judgements rather than following rules, they're the ones most likely to drift toward optimising for narrow metrics (speed, cost, volume) while losing sight of what the organisation actually cares about (fairness, quality, customer trust). These agents need clear boundaries and active human oversight.

Finally: multi-agent orchestration

One agent directing a network of sub-agents. This is where governance becomes genuinely difficult. When agents supervise other agents, which governance standard applies? If a trusted orchestrator delegates to a less-tested sub-agent, who is accountable? These are questions the industry is still working through, and where governance approaches will need to keep pace with the technology.

At SEEK we have focused on getting the fundamentals right at the earlier stages before moving to more advanced agents. While there may be temptation to jump to sophisticated multi-agent deployments, particularly when competitors appear to be moving fast, resisting that temptation until simpler deployments are operating reliably is likely to be an important governance decision.



Data governance: agents are not human users

When a human accesses organisational data, professional judgement and personal accountability shape how it's used. While agents can be given rules and constraints that reflect professional standards, they don't exercise judgement the way a human does, and they can't be held personally accountable; yet they can access data at machine speed, across multiple systems at once.

Treating agents as digital equivalents of human users, and giving them the same access permissions, comes with governance risk.

We think agents should operate under a minimum viable data access model: the narrowest possible permissions needed to complete a defined task, revoked immediately when the task is done. Scope creep in data access isn't just inefficient; it increases the risk of data exposure and unauthorised action.

We're also exploring the idea that AI-generated content should carry a provenance label, marking it as AI-produced, alongside a reliability classification showing whether it's been human-verified, partially reviewed, or passed through unreviewed. Over time, these labels give leadership an honest picture of how dependent decision-making has become on autonomous outputs.

Democratising access: the tension between speed and control

One of the harder questions we've grappled with is who gets to build and deploy agents within an organisation.

Restricting agentic AI to specialist technical teams is safe but slow. It limits the organisation's ability to discover where agents can add value, and it concentrates AI capability in a small group rather than building broad organisational literacy.

On the other hand, giving everyone open access without guardrails creates exactly the shadow AI risk that governance frameworks are designed to prevent: agents operating outside oversight, accessing data they shouldn't, and producing outputs without human oversight and accountability.

We've been experimenting with a middle path: a governed self-service platform where any staff member can create and use agents within a deliberately constrained environment. The platform respects existing access permissions, so an agent built by a staff member can only access what that person is already authorised to see. Staff start with minimal permissions and can expand their capability through a structured assessment as they demonstrate understanding.

Alongside the platform, we've invested in AI literacy training for all staff, covering not just technical skills but governance awareness and the ability to judge when an agent needs supervision versus when it can be trusted to operate.

Early signs are encouraging. Staff are discovering automation opportunities that no central team would have identified, while shadow AI risk appears contained because the sanctioned environment is easier to access and more capable than unsanctioned external tools. However, we expect our approach to continue evolving as we learn more about what works at scale.

Where governance needs to keep evolving

Agentic AI is developing quickly. Any governance approach worth adopting needs to be designed for change. Based on our experience, we see several areas where the industry's collective thinking still needs to mature.

Governing multi-agent interactions

When multiple agents interact, they can produce outcomes that weren't anticipated when any individual agent was designed. As multi-agent systems become more common, governance approaches will need to move beyond assessing agents individually toward understanding system-level behaviour.

Keeping human oversight meaningful

As agents become more capable, the risk of automation bias grows. Supervisors need the right tools, training, and authority to maintain genuine oversight and accountability rather than becoming a rubber stamp. This requires ongoing investment and operating model design, not a one-off training programme.

Balancing governance cost with business value

Every governance measure has a cost. Organisations that over-govern will struggle to realise the benefits of agentic AI. Organisations that under-govern will carry unnecessary risk. Getting this balance right requires honest assessment of each use case, and a willingness to adjust as the risk profile changes.

Preparing for agent failures

As the OpenClaw incident described earlier in this paper showed, even well-designed guardrails can fail. Organisations need incident response plans that are specific to agentic AI, not just repurposed from existing IT playbooks.

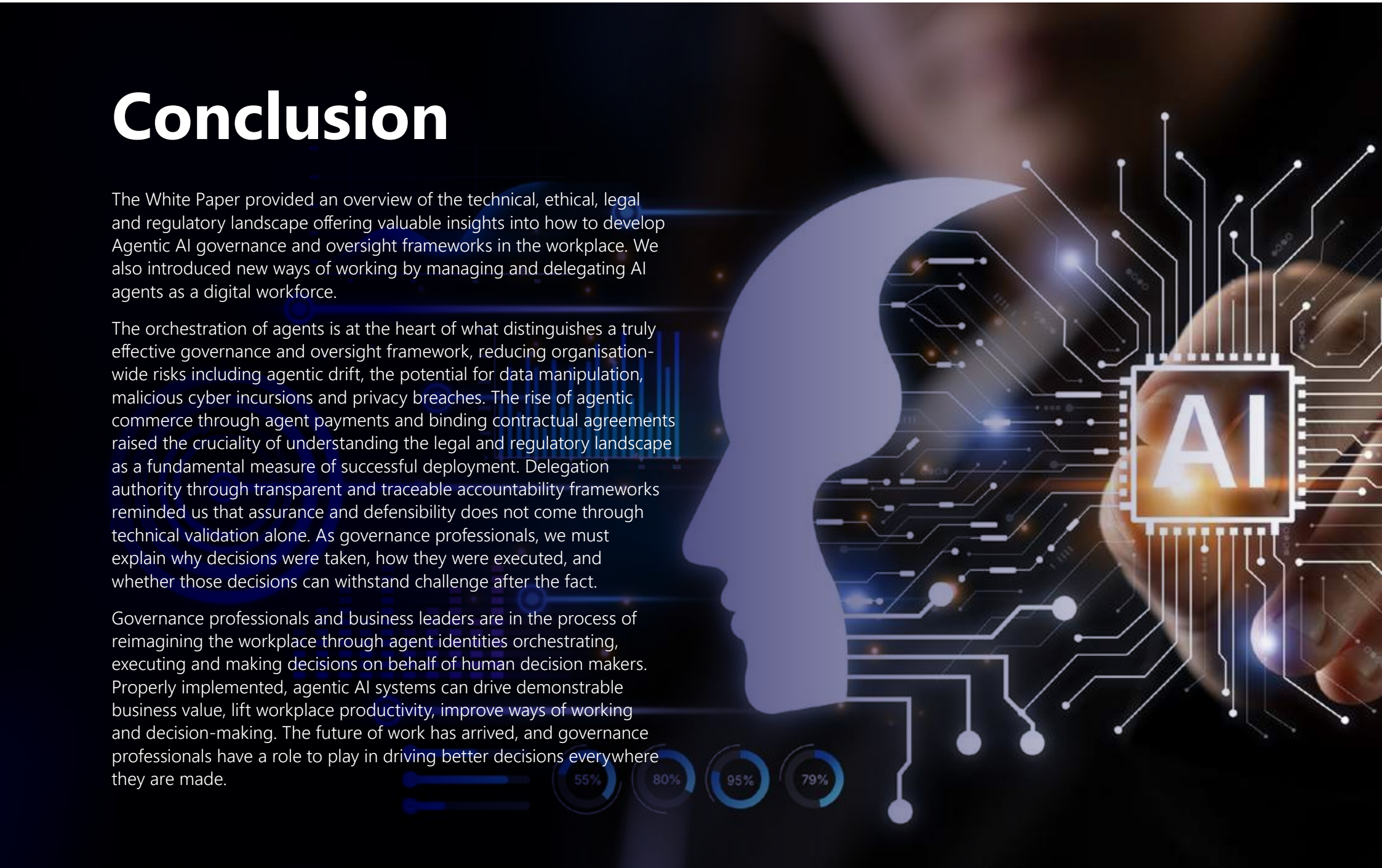
These are not reasons to slow down; they are reasons to build practical governance that can adapt. The organisations that get the most value from agentic AI will be those that treat governance as a living practice.

Conclusion

The White Paper provided an overview of the technical, ethical, legal and regulatory landscape offering valuable insights into how to develop Agentic AI governance and oversight frameworks in the workplace. We also introduced new ways of working by managing and delegating AI agents as a digital workforce.

The orchestration of agents is at the heart of what distinguishes a truly effective governance and oversight framework, reducing organisation-wide risks including agentic drift, the potential for data manipulation, malicious cyber incursions and privacy breaches. The rise of agentic commerce through agent payments and binding contractual agreements raised the cruciality of understanding the legal and regulatory landscape as a fundamental measure of successful deployment. Delegation authority through transparent and traceable accountability frameworks reminded us that assurance and defensibility does not come through technical validation alone. As governance professionals, we must explain why decisions were taken, how they were executed, and whether those decisions can withstand challenge after the fact.

Governance professionals and business leaders are in the process of reimagining the workplace through agent identities orchestrating, executing and making decisions on behalf of human decision makers. Properly implemented, agentic AI systems can drive demonstrable business value, lift workplace productivity, improve ways of working and decision-making. The future of work has arrived, and governance professionals have a role to play in driving better decisions everywhere they are made.



List of references

1. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/reimagining-the-value-proposition-of-tech-services-for-agentic-ai
2. https://www.cyberdaily.au/tech/11928-more-than-half-of-australian-businesses-are-already-using-agentic-ai
3. https://cfotech.com.au/story/australian-firms-to-boost-agentic-ai-investment-to-usd-12-3-billion
4. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027
5. https://xmpro.com/agentic-ai-for-industry-cutting-through-the-hype-to-unlock-real-industrial-value/#:~:text=Avoiding%20%22Agent%20Washing%22,it%20provide%20unique%2C%20transformational%20value?
6. https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf
7. https://www.mallesons.com/au/en/insights/latest-thinking/agentic-ai-rogue-agents-real-liability.html?redirect
8. https://mitsloan.mit.edu/ideas-made-to-matter/agentic-ai-explained
9. https://www.deloitte.com/au/en/services/consulting/perspectives/human-agentic-workforce.html
10. https://itbrief.com.au/story/ai-agents-surge-in-australia-but-integration-lagging
11. https://medium.com/@community_md101/how-data-governance-improves-ai-success-65cc4feb3fbc
12. https://www.mallesons.com/au/en/insights/latest-thinking/agentic-ai-rogue-agents-real-liability.html
13. https://mashable.com/article/what-is-clawdbot-how-to-try
14. https://techcrunch.com/2026/02/23/a-meta-ai-security-researcher-said-an-openclaw-agent-ran-amok-on-her-inbox/; https://x.com/summeryue0/status/2025774069124399363
15. https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents
16. M. Galster, S. Mohsenimofidi, J.L. Lulla, M.A. Abubakar, C. Treude, & S. Baltes (2026). Configuring Agentic AI Coding Tools: An Exploratory Study. arXiv:2602.14690. https://arxiv.org/pdf/2602.14690v1
17. Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society, 5, 40–60. https://doi.org/10.17351/ests2019.260
18. https://mashable.com/article/moltbook-security-risks
19. https://www.securityweek.com/security-analysis-of-moltbook-agent-network-bot-to-bot-prompt-injection-and-data-leaks/
20. Electronic Transactions Act 1999 (Cth) s 15C.
21. Jeannie Marie Paterson and Elise Bant, 'The Undue Influence of AI Assistants, Agents and Companions' (2026) 19 *Journal of Equity* 107
22. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5028371
23. *Amazon.com Services LLC v Perplexity AI, Inc* (ND Cal, No 25-cv-09514-MMC, 9 March 2026)
24. Australian Competition and Consumer Commission, *Recent Developments in Artificial Intelligence: Industry Snapshot* (December 2025) 22
25. Article 22 of the General Data Protection Regulation in the European Union
26. *Thaler v Commissioner of Patents* [2021] FCA 879, [8]; *Thaler v Comptroller-General of Patents, Designs and Trade Marks* [2023] UKSC 49, [75]
27. *Moffatt v Air Canada* [2024] BCCRT 149
28. The Treasury (Cth), *Review of AI and the Australian Consumer Law* (Final Report, October 2025) 24-25.
29. *Australian Securities and Investments Commission, Beware the gap: Governance arrangements in the face of AI innovation* (Report 798, October 2024) 34.
30. *Australian Securities and Investments Commission, Corporate Plan 2025–26* (2025) 11.
31. Corporations Act 2001 (Cth) s 912A
32. *Australian Securities and Investments Commission v RI Advice Group Pty Ltd* [2022] FCA 496.
33. Nada Madkour et al, *Agentic AI Risk-Management Standards Profile* (Report, UC Berkeley Center for Long-Term Cybersecurity, February 2026) 47; Infocomm Media Development Authority, *Model AI Governance Framework for Agentic AI* (January 2026) 13, 16-17.
34. Digital NSW, *AI Agent Usage and Deployment Guidance* (Guidance Document, October 2025) 4.
35. World Economic Forum, *AI Agents in Action: Foundations for Evaluation and Governance* (White Paper, November 2025).
36. Madkour et al (n 14).
37. IMDA (n 14) 11; WEF (n 17) 25-26; Madkour et al (n 14) 36.
38. IMDA (n 14) 4-5; see also WEF (n 17) 13-14; Madkour et al (n 14) 17.
39. IMDA (n 14) 16-17; Madkour et al (n 14) 36; WEF (n 17) 26.
40. IMDA (n 14) 11-12; WEF (n 17) 26.
41. IMDA (n 14) 19-20; WEF (n 17) 18-19.
42. Madkour et al (n 14) 35, 41.
43. IMDA (n 14) 19; Madkour et al (n 14) 41-42.
44. Digital NSW (n 15) 7-8; WEF (n 17) 6, 27; IMDA (n 14) 18.
45. IMDA (n 14) 13, 15; Madkour et al (n 14) 8; Digital NSW (n 15) 4.

Appendix – Agentic AI readiness checklist

Technical capabilities

- ◻ Understand the technical characteristics and different types of AI agents.
- ◻ Understand how AI agents make decisions, how they are trained, their inherent capabilities, how agents learn and execute autonomous activities
- ◻ Determine the value proposition of AI agent systems to orchestrate complex workflows by assessing agent risk profiles and the tangible benefits of AI agent systems
- ◻ Develop an agentic AI risk matrix
- ◻ Develop a risk tolerance statement and assess and adjust in line with the organisation's risk appetite for using agentic AI systems.
- ◻ Consider whether the AI agent can effectively leverage high-quality data to learn, reason and act on decisions.
- ◻ Ensure the organization has a data governance framework in place that prioritises data quality and integrity.

Ethical considerations

- ◻ Identify and communicate the difference between GenAI governance frameworks and Agentic AI frameworks by contemplating the issues of:
- ◻ delegation authority
- ◻ traceable accountability
- ◻ continuous monitoring and assessment
- ◻ ethical defensibility.
- ◻ Consider how safeguards such as 'kill switches' and 'dead man's switches' are effectively deployed and executed when an agent goes rogue?
- ◻ Is one legally or morally culpable if an agent goes rogue despite all best efforts put in place, including guardrails such as commands in agents?
- ◻ Identify the ethical principles and guidelines that apply to the different stakeholders of agentic AI, who have different stakes including different levels of responsibility and culpability when things go wrong?
- ◻ Consider how to rebuild trust where it is violated between humans and the AI agent and between the different human stakeholders?
- ◻ Consider use-cases where things have gone wrong and how the organisation can learn from past behaviours and mistakes.
- ◻ Contemplate ethical dilemmas arising from emergent or unforeseeable risks.
- ◻ Consider the benefits of open-agent models that may learn, reason and act on ungoverned or low-quality data.

Legal and regulatory landscape

- ◻ Map the key legal risks the deployment of AI agents introduces.
- ◻ What are the key legal risks associated with:
- ◻ Data security and information privacy
- ◻ Entering into contracts
- ◻ Contracts have been breached
- ◻ Existing regulatory requirements

- Negligence, risk of property damage or personal injury
- External communications, evidence and legal professional privilege
- Understand what legal risks might the use of an AI agent amplify.
- Understand the potential scope of an organisation's liability for its AI agents
- Understand why existing AI governance needs to evolve to deal with the risks of agentic AI systems.
- Effectively communicate:
- The evolution of agentic AI systems
- Agentic AI systems inherent risks
- The need for an Agentic AI governance and oversight framework to effectively, securely and ethically manage agentic AI systems.

Agentic AI governance and oversight

- Develop an understanding of the new governance measures organisations should consider implementing for agentic AI.
- Consider governance accountability arrangements pre-deployment, during testing and piloting, during deployment and broader governance considerations.
- Map the AI ecosystem and identify where risks arise across the agentic supply chain including:
  - model provides
  - system providers
  - tools and integration
  - deploying organisations
  - internal users and
  - external users.

Preparing directors and Board meetings

- Identify how a director can demonstrate that an AI agent's decisions and actions were reasonable, lawful and aligned with organizational intent.
- Prepare and document delegation boundaries?
- Prepare and document structured impact assessment prior to deployment and updated over time
- Ensure decision traceability and human oversight at the systems level
- Codify duty-based execution controls
- Ensure independent assurance and audit mechanisms are in place
- Prepare for objection and remediation
- Identify how directors can test and verify the integrity of agentic AI systems
- Understand that sole reliance on technical validation to verify the integrity of agentic AI systems is insufficient
- For operational integrity ensure the organization possesses explicit delegation structures
- Does the organization have structured impact assessments?
- Do decision logs enable the reconstruction of inputs, outputs, model versions, data sources and intervention points?
- Assess distributive effects, including whether outcomes disproportionately burden vulnerable groups or diminish dignity and agency
- Verification processes should include assessment of likely public scrutiny, clarity of stakeholder communication, and readiness for rapid remediation where harm or controversy arises

Questions for the Board

Understanding exposure

- ☐ What AI agents are being used throughout our organisation?
- ☐ Do we have third-party software that might include AI agents? What about shadow use of AI agents?
- ☐ Do we properly understand the key risks associated with our use of AI agents?
- ☐ What risks do we face if an agent makes a significant error, such as making a misleading statement to a customer, entering into an unauthorised transaction, or breaching privacy or data security obligations?
- ☐ Have we properly assessed our cybersecurity risks arising from our deployment of AI agents?
- ☐ Do our existing insurance arrangements cover us for potential AI agent-related claims? Do our vendor contracts ensure that liability for AI agent-related risks is appropriately allocated?

Implementing AI governance

- ☐ Does our AI governance framework specifically address the risks of agentic AI, or is it designed for simpler AI use cases?
- ☐ Does our data governance framework address how AI agents use, process and share the data they access?
- ☐ Does it deal with the quality and integrity of data being accessed by AI agents?
- ☐ What guardrails have we established around our deployment of AI agents to ensure that they only operate within our risk appetite?
- ☐ Does our AI governance framework identify who is accountable for each AI agent used within our organisation? Is that accountability clearly understood?
- ☐ Are we confident that mitigations such as human oversight are being implemented in practice in an appropriate and meaningful way (and not just as a formality)?

Assurance and capacity

- ☐ Are we appropriately testing and assuring AI agents in safe environments before we deploy them?
- ☐ Are we testing and monitoring agentic systems on an ongoing basis post-deployment?
- ☐ Are we applying an appropriate level of diligence to third-party AI tools and integrations?
- ☐ To what extent are we relying on vendors' claims about security and other agentic risks without checking them?
- ☐ Do we have processes and systems in place to ensure that our personnel are being appropriately trained to work with and oversee these systems effectively?
- ☐ Do we have the expertise needed to test and maintain these agentic AI systems and to deal with incidents if they fail?

Governing AI agents as a digital workforce

- ☐ **Acknowledge that deploying an AI agent is an act of delegation**
  - ☐ Ensure clear authority, defined boundaries and responsibilities for AI agents and staff managing them.
  - ☐ Test how safely each agent works alongside other functions across the organisation.
  - ☐ Consider worker strategies to monitor the safety and veracity of agent-to-agent interactions.

- **Understand the evolving role of human oversight**
  - Consider investing in dedicated oversight tools such as dashboards that give supervisors real-time visibility into agent performance, anomaly flags and cost, rather than periodic review.
  - Consider governance frameworks that move beyond assessing agents individually toward understanding system-level behaviour.
  - Consider supervisor training and tools required to maintain genuine oversight and accountability.
  - Develop incident response plans that are specific to agentic AI.
- **Consider maturing progression of AI agents**
  - Start with smart automation, human-AI collaboration, contextual intelligence and multi-agent orchestration.
  - Provide agents with access to minimum viable data with the narrowest possible permissions needed to complete a defined task.
- **Decide who gets to build and deploy AI agents across the organisation**
  - Consider the benefit of any staff members creating and using agents within a constrained environment.

Governance Institute of Australia

Another Good Decision

Governance Institute of Australia
1800 251 849
Level 11/10 Carrington Street,
Sydney NSW 2000
www.governanceinstitute.com.au