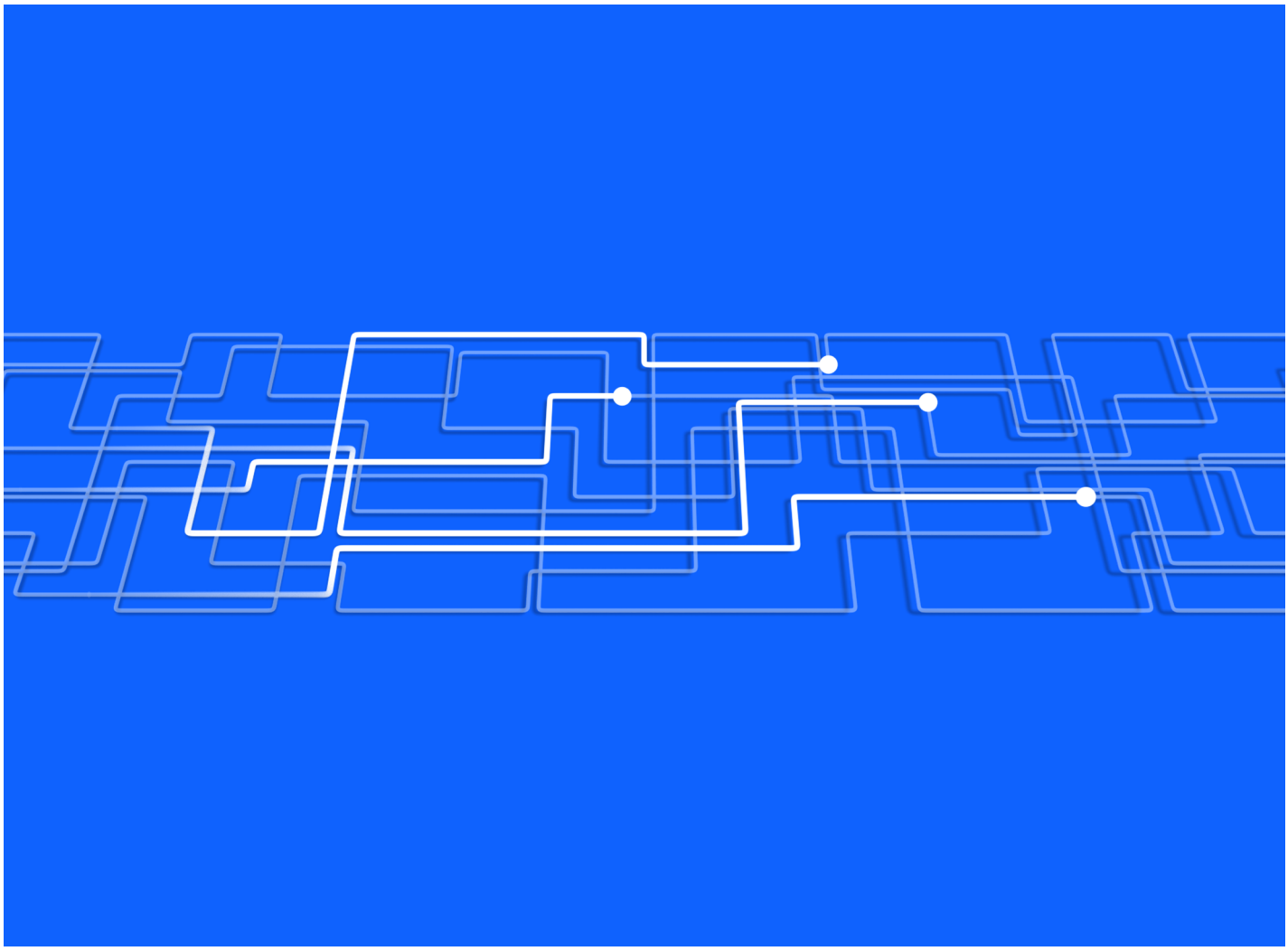
IBM Consulting



Agentic AI in Financial Services

Opportunities, Risks, and Responsible Implementation

IBM

Table of contents

Executive Summary

The advent of Agentic AI represents a significant milestone in the ongoing AI super cycle, characterised by rapid technological advancements and substantial investments. Agentic AI has emerged as a highly usable and dynamic technology early in its development, presenting unique opportunities and challenges. This paper explores the implications of Agentic AI in the financial services sector, emphasising the transformative potential and the necessity for robust risk management strategies.

The AI Super Cycle

The global economy is currently experiencing an AI super cycle, driven by unprecedented progress and investment in AI technologies. This cycle is igniting business transformation initiatives aimed at accelerating growth and uncovering new efficiencies. Tech vendors are competing to establish their platforms as the leading AI environments, while companies strive to integrate AI into their products and services to stay ahead of the competition and enhance customer engagement.

The rapid adoption of AI is shifting the focus from traditional business applications to data fabric and AI, creating a new software arms race. Control over AI models, user interfaces and data integration are becoming crucial. AI's early usability and dynamic nature offer the potential to address previously complex business problems, accelerating the codification of business processes and converging various risks, thereby transforming human-technology interactions.

The Tipping Point

Despite significant investments in AI, many initiatives have failed to realise substantial business value. The true value of AI will be achieved when humans can completely delegate tasks to AI systems, both simple and more complex tasks which can often result in poor customer/user outcomes, allowing humans to focus on more strategic and higher value activities. This transition will mark a tipping point in AI value realisation. With that said, human oversight remains an important element, even if humans delegate more tasks to the agent/system. As such, it is vital to consider what measures organisation could take as part of their AI Governance approach, to unlock this value while not exposing the organisation to unintentional risk.

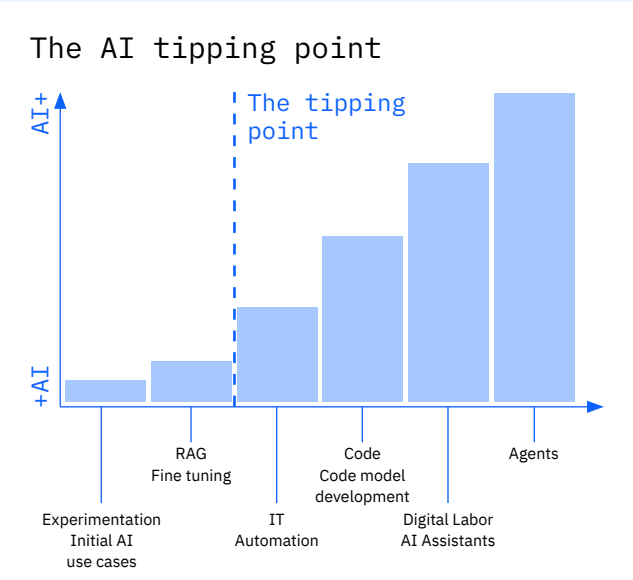


Figure 1: The AI tipping point, from which, the return on investment on AI is believed to realise.

Impact on Risk Management: Trust and Governance

Agentic AI introduces a novel risk landscape, requiring a shift in risk management practices. The autonomous nature of AI agents complicates human oversight, making real-time intervention challenging. This necessitates a proactive approach to risk management, integrating AI systems with robust guardrails and real-time monitoring to ensure safety and reliability.

Core to ongoing governance is the need to develop robust registration that maps out usage of AI agents across use cases and relying on this to determine impacts to material change and performance issues.

Establishing trust in AI agents is crucial. This involves implementing standardised guardrails as modular, reusable components across different use cases, and ensuring real-time monitoring of AI actions. The paper advocates for a "compliance by design" mindset, where risk mitigation strategies are developed alongside AI systems, ensuring alignment with organisational risk appetite and validating use cases before significant investments.

Successful management of Agentic AI requires collaboration across various organisational roles, from HR and change management to data scientists and security analysts. Clear governance structures and communication pipelines are essential to navigate the new and amplified risks associated with AI agents.

Practical Steps for Implementation

Organisations should adopt a strategic, phased approach to incorporating Agentic AI. This involves identifying business value for potential use cases, defining detailed personas and goals, defining risk appetites, updating risk assessment processes, and implementing controls to manage AI-specific risks effectively. Starting small and refining the approach will ensure scalability and trustworthy implementations.

Conclusion

Agentic AI presents exciting opportunities and unique challenges for the financial services sector. By understanding the capabilities and limitations of this advanced technology, financial institutions can harness their potential while mitigating associated risks. Through strategic planning, robust risk management, clear control and supervision requirements and a commitment to responsible AI practices, the financial sector can successfully navigate this new frontier. Future papers will explore the broader impacts of Agentic AI on skills, culture, technology architectures, and customer and employee engagement.

Introduction

What are AI Agents

An agent, in the context of artificial intelligence (AI), is a software entity capable of perceiving its environment and acting upon it with a high degree of autonomy. It is a system designed to reason through complex problems, interpret and create actionable plans, and execute these plans using a suite of tools.

The evolution of "Agents"

Agents have been leveraged for as long as we have leveraged computers. In robotic process automation (RPA) a software robot acts as an agent for a user or application to run specific tasks. These rule-based agents typically leverage a pre-determined input and output; with simple logic in code to determine the course of action.[1]

Over the years, natural language processing and understanding (NLP and NLU) methodologies evolved as traditional AI capabilities; and we have come to use these capabilities as virtual agents in the form of AI chatbots. These chatbots, as opposed to automation, do not require explicit inputs. Rather, they are able to classify intents. Meaning, given user input, they attempt to associate it with a known course of action, for which they have a pre-configured, deterministic logic or "a user journey". These agents then respond via a pre-determined output.

With the introduction of Large Language Models (LLMs) and generative AI, this natural language processing capability has improved significantly; enabling better intent classification and generation of unique outputs. These LLMs produce outputs that are not pre-determined but rather they determine responses based on learned patterns from large datasets they have been trained on, and user prompts. Many current chatbots and implementations of generative AI leverage this capability and extend the models knowledge base, beyond the data it is trained on, to specific data sources to improve outcomes. This is achieved in a largely deterministic manner, with controlled parameters and retrieval methods, through methods such as Retrieval Augmented Generation (RAG).[2]

Emerging AI agents, employing LLMs, go beyond the capabilities introduced by LLM-based chatbots. AI agents can autonomously, through its design, perform a wide range of functionalities beyond natural language processing including decision-making, problem-solving and interacting with external environments via defined tools.[3] They are characterised by 'under specification', the ability to accomplish a goal provided by a user without a concrete specification of how the goal is to be accomplished; and 'long-term planning', the ability to reason and make interim decisions and predictions that will affect the next action they perform to achieve their goal.[4] This drives an adaptability rendering AI agents particularly suitable for intricate tasks in dynamic environments such as financial services.

Core components of an AI Agent

AI agents have a notion of planning, reflection and other control structures that heavily leverage the model's inherent reasoning capabilities to accomplish a task end-to-end. Reasoning is a fundamental building block enabling AI agents to solve complex problems effectively. Combined with the ability to interact with the external environment, through tools, AI agents are empowered to execute more general-purpose work.[5]

1 IBM – The evolving ethics and governance landscape of agentic AI
2 What is retrieval-augmented generation (RAG)? - IBM Research
3 IBM – What are AI Agents
4 2302.10329 - Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., ... & Maharaj, T. (2023, June) -et al., Harms from increasingly agentic algorithmic systems., 2003
5 The Landscape of Emerging AI Agent Architectures for Reasoning, Planning and Tool Calling

Model:
The LLM that will be the centralised operations engine for agentic tasks. The model drives how the agent understands and responds to inputs, events and conditions and dictates its behaviour.

Tools:
The mechanism that uplifts the agent's native abilities. These tools can take a range of forms such as functions, operations to manipulate data stores, or API calls. Tools may be selected during the execution as opposed to deterministically in advance, based on the context of the tasks/subtasks planned.

Reasoning & Planning Layer:
The model's cyclical process of goal initialisation, planning, reasoning, action and reflection. It enables the autonomy by which the AI agent, given its tools, manages its internal state and interacts with its environment to process information and determine action, until it achieves its goal or a stopping point. This can have greatly varied complexity depending on the task. This layer inclusive of:

- **Goals, Instructions & Personas:** The defined or instructed outcomes, or motivations by which, the agent is working to achieve, as well as the conditions under which it may stop even without obtaining its goal. This can be supported by defined personas/roles to further direct toward meaningful outcomes. Personas also contain descriptions of the tools they have access to.
- **Memory:** Sophisticated short- and long-term memory features allow agents to retain and utilise information across interactions, as well as enabling coherence within a single interaction.
- **Planning, Reasoning & Critiquing:** Reasoning frameworks are leveraged as building blocks to enable the cyclical problem-solving process and guide the AI agent, leveraging its model, in interacting with its environment to achieve a goal. Some common reasoning frameworks leveraged to enable this are Chain-of-Thought (CoT),[6] Reasoning & Act (ReAct)[7] and Reasoning Without Observation (ReWOO).[8]

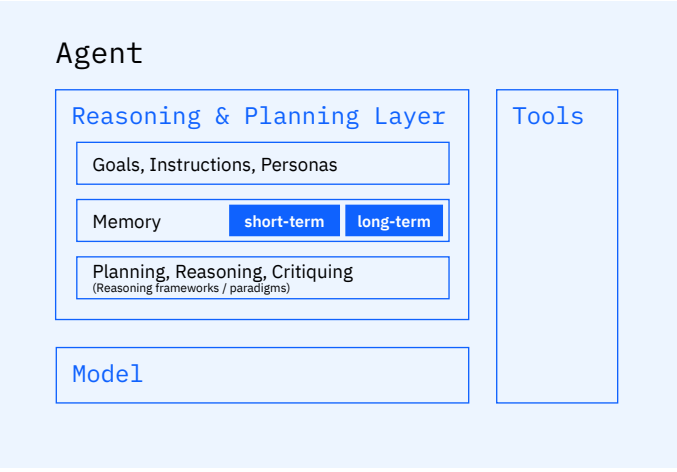


Figure 2: The components of an AI Agent. Diagram altered from Google AI Agents.[9]

Agentic AI Systems and Orchestrations

Orchestration, in the context of computer systems, is the automated coordination of and management of applications, services and data. **AI systems** are orchestrations that include at least one AI model.[10] Effective AI orchestration streamlines the end-to-end AI lifecycle and enables greater efficiency, scalability, responsiveness and effectiveness.[11]

Agentic AI systems are characterised by orchestration where one, or more, AI agents are involved in complex problem solving. Whilst AI agents are singular models with native tools that can autonomously plan and execute tasks, an agentic system includes orchestration that enables integration of an agent with other agents, models, tools and data sources. Agentic AI systems which include multiple AI agents (multi-agent systems) are of special interest. While the AI agents in the system may collaborate in tackling complex problems, each could still have its own goals, tools, and capabilities.

In considering the implementation and orchestration of multi-agent systems, depending on the design of the system, users are broadly interacting with three types of AI agents:

- **Principal Agents:** Understand the objective and dynamically plan and orchestrate with other agents and services, to achieve an outcome.
- **Service Agents:** Experts with fit for purpose tools to drive specialised domain knowledge or expertise and drive execution against a plan.
- **Task Agents:** Micro-operators with limited knowledge boundaries specialised for execution on fine-grain tasks.

These agents aggregate different approaches to agent building from goal-based agents to simple reflex agents that operate strictly within a defined boundary.[12] Note, the design of your system will determine how the agents interact with users, customers or their tools. By example, to execute against a goal, a Principal agent might coordinate with Service agents who in turn access their tools, further Task agents, and human verification.

[6] IBM – What is Chain of Thoughts (COT)?
[7] What is a ReAct Agent? | IBM
[8] What Is Agentic Reasoning? | IBM
[9] Google - Agents
[10] IBM – What is AI Orchestration?
[11] IBM – What are Compound AI Systems?
[12] AWS – What Are AI Agents

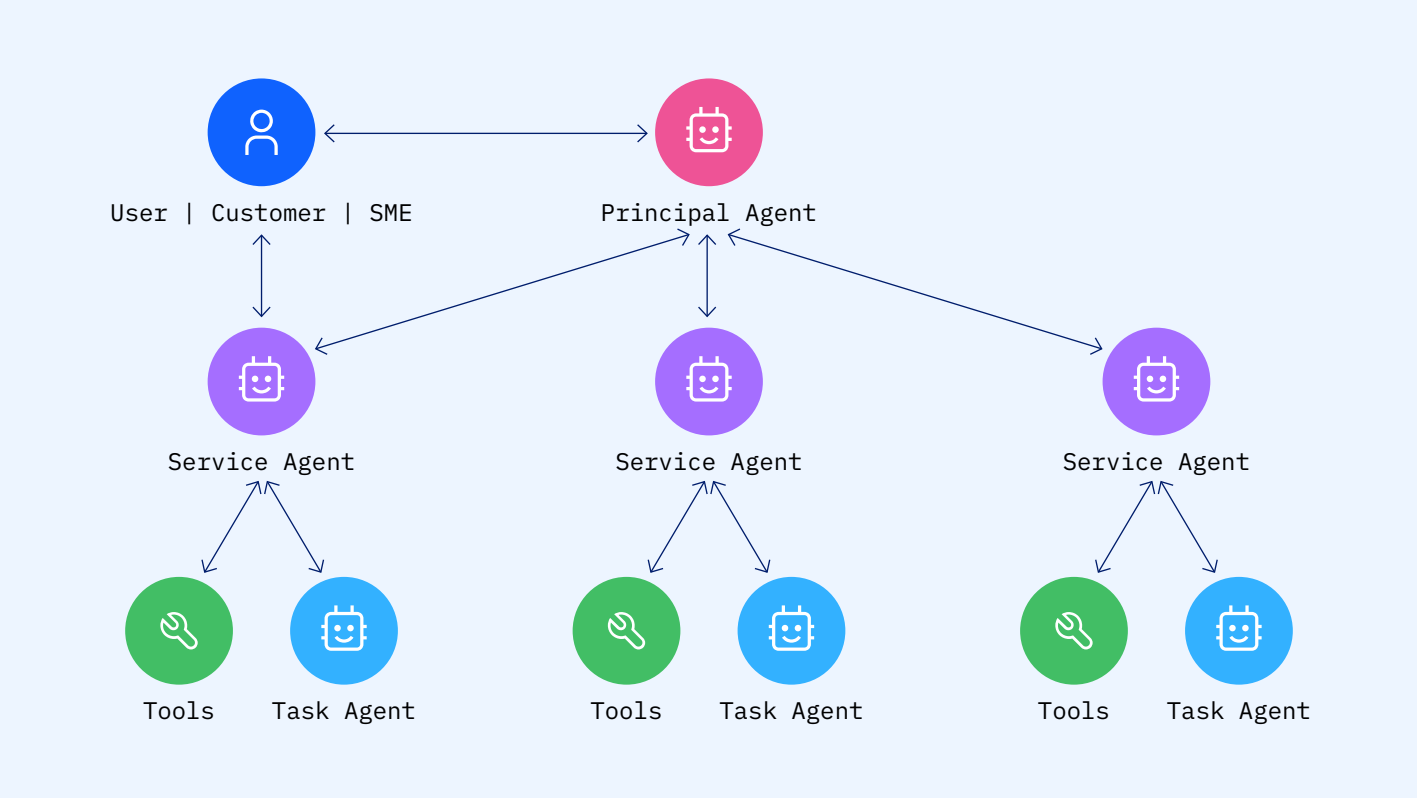


Figure 3: An example of a multi-agent AI system which includes a Principal agent, Service agents, and Task agents.

The orchestration of these AI agents, and collaboration in Agentic AI systems, drives significant adaptability to a changing business environment. It enables autonomous coordination of action beyond a single function, proactive processes and insights, dynamic event-driven actions, leveraging of domain-specific knowledge and rapid change in response to new information. The degree of autonomy employed in an AI system's orchestration is an additional differentiator of Agentic AI systems. Lower levels of autonomy would leverage human-defined, fully programmed, control logic to orchestrate between the components of the system.

Increasingly agentic AI systems leverage an orchestration of the system led by a principal agent, or simply an LLM, to dynamically direct the workflow and operations of the various components of the system to achieve an outcome. These operations are inclusive of, by example; data management and preprocessing, coordinating other agents, managing LLM resources and performance, prompt chain management, API interactions and state management.

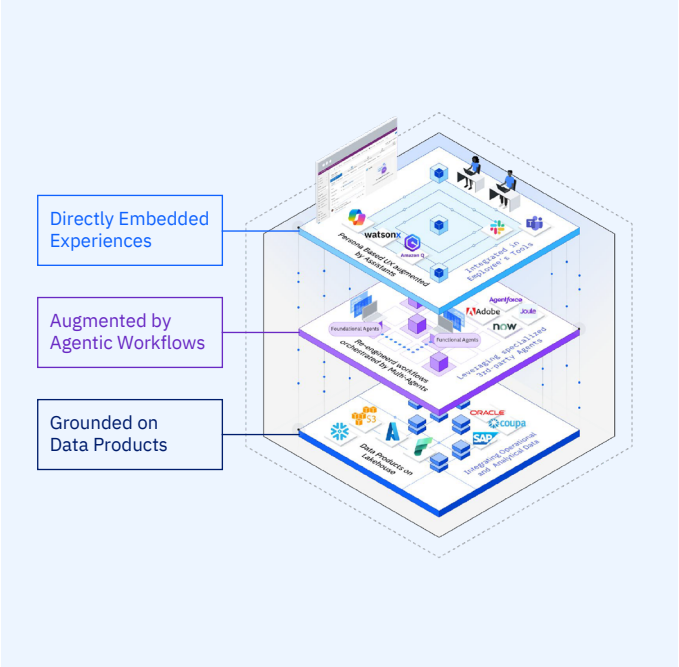
For well-defined problems, programmatic orchestration of LLMs will likely create more efficient workflows and predictable outcomes. However, for systems designed to handle a variety of complex queries, an agentic approach, leveraging one or many agents with specialised skills, allows flexibility, adaptability and reduction in the effort required to define increasingly complex business processes.

Opportunities in the Financial Services

The financial services industry stands to reap substantial benefits from AI agents. There has long been a reliance on rigid legacy systems to structure business processes which have proven to be obstacles to efficiency and agility. AI agents represent an opportunity to shift the approach to building technology services from responsive to adaptive. Where today new technology services are largely built as a response to a changing business environment, AI agents allow for the building of technology services that dynamically respond and adapt. This can empower more accessible, personalised, banking services and experiences for our customers. Previously complicated workflows would require multiple human interactions, legacy system processes and operational team handoffs. AI agents, and agentic systems, allow for entirely new customer experiences and outcomes that limit the friction and complexity of these interactions.

In the practical realisation of these new experiences, leveraging the components and building blocks of AI agents, every agentic system:

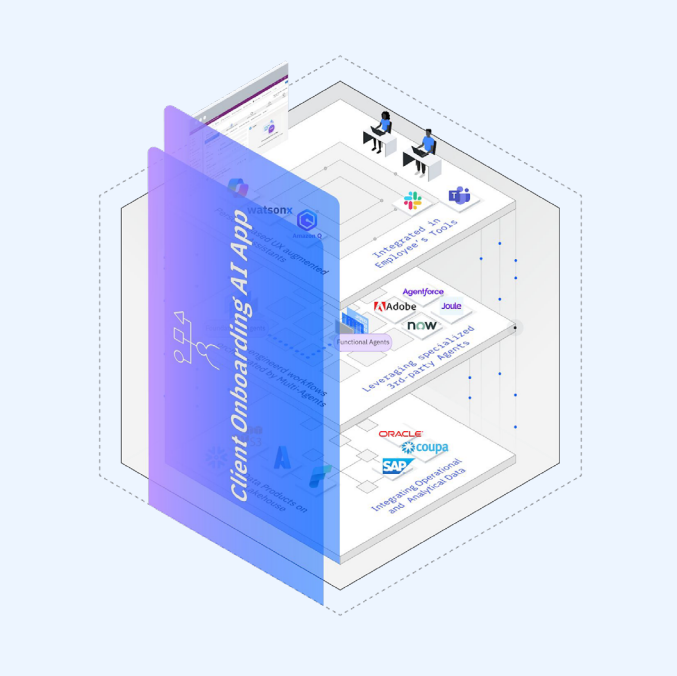
- Seeks to transition from multiple, non-integrated interfaces to consistent, persona-based experiences embedded directly into the tools used by our users.
- Is supported by the development, and orchestration of, specialised service and task agents who enable specific business processes, understand specific domains, or operate seamlessly across comprehensive business environments
- Is grounded on trusted, well-defined, data products delivering high quality operational and analytical insights to seamlessly support the end-to-end business processes
- Is enabled by robust data governance frameworks to maintain accuracy, security, and compliance across all data sources



Let's consider this through the lens of a typical business process, and challenge, in Financial Services like customer onboarding. How the customer, our customer service representatives and the supporting agentic system interact will differ depending on the design of the system and its orchestration. This orchestration may be designed leveraging an architectural framework, such as LangGraph,[13] to coordinate the workflows within the system. It can be either deterministic or directed by an agent or LLM, depending on desired outcomes.

In a customer onboarding journey using an agentic system, a customer may interact directly with a customer service representative, who then interacts with the system to complete the onboarding process. Through this design, the interactions with a customer will always be directed by a human representative who returns responses based on the supported outputs of the agentic system. The principal agent of the agentic system dictates the steps, based on the business process, to be executed to complete onboarding.

[13] LangGraph



By example the onboarding and KYC business process, leveraging an agentic system, may function in the following way:

1. A customer applies for account opening and provides supporting documents and information.
2. A customer service representative coordinates with, and is supported by, the agentic system to complete this onboarding.
3. A principal agent dictates the thought, reasoning and action by which the business process is executed. It leverages the orchestration framework of its designed agentic system to determine the boundaries of its capabilities and its access to the business environment.
4. The principal agent directs actions to domain specialised service agents; for example, a risk analysis agent responsible for evaluating a customer's risk profile and a sanctions agent responsible for screening customer data.
5. The service agents then interact with, and direct tasks to, task agents such as document validation agents who verify completeness and accuracy and a compliance monitoring agent who verifies defined AML/KYC regulations are met as business processes are executed. Additionally, they might interact with task specialised customer due diligence and enhanced due diligence agents.
6. The service and task agents, via the principal agent, may dictate human-in-the-loop action where high-risk customers are identified or the need for additional documentation.
7. The customer service representative ensures and verifies the outcome.

An agentic system in this example enables extensive data collection, verification and processing across disparate systems. This greatly enhances the customer experience through execution across an often highly fragmented and manual business process.

This is illustrative of the significant opportunity across all aspects of the value chain. For example, automating routine tasks, such as data entry and basic customer service inquiries, could improve operational efficiency and reduce human error. AI agents could enhance risk assessments, provide personalised recommendations based on individual client profiles, and strengthen fraud detection by identifying complex patterns indicative of fraudulent activity. The potential for increased efficiency, improved customer experience, and better decision-making is considerable.[14]

There are 3 core emerging patterns of implementation in Financial Services with high potential, and associated business value, for application of Agentic AI systems:

- AI-Powered Customer Engagement & Personalisation
- AI-Driven Operational Excellence & Governance
- AI-Augmented Technology & Software Development

Across these categorisations, the below is representative of emerging areas to drive business value through the development of Agentic systems.

Case Study Pattern: AI-Powered Customer Engagement & Personalisation

There is significant opportunity for enhancing customer experiences and enriching customer interactions in financial services. Typical emerging applications focus on, but are not limited to:

- Customer Service & Engagement
- Hyper-personalisation
  - Product and service offer customisation
  - Dynamic pricing and deal optimisation
  - Recommendation & Robo-advice (e.g. ensuring appropriate products and services)
  - Behaviour and preference driven interactions (e.g. loyalty offers)
- Onboarding, KYC & AML Optimisation

[14] https://www.akira.ai/blog/risk-management-with-agentic-ai

Case Study Pattern:
AI-Driven Operational Excellence & Governance

There is significant opportunity to optimise middle/back-office operations to reduce risk, enhance compliance, streamline workflows and administrative overhead and drive better business and customer outcomes. Typical emerging applications focus on, but are not limited to:

- Lending & Loan Approvals
- Account Operations
  - Transfer ownership of assets
  - Power of Attorney
  - Account freezing
- Anomaly Detection
  - Transaction Monitoring
  - Fraud Detection
- Automated Risk Management & Compliance
  - Control effectiveness verification (requirements against implementation)
  - Execution of control operations
  - Execution of control performance and continuous monitoring
  - Product/service compliance: by example leveraging, models as a judge to verify quality of outcome for the customer and that compliance obligations were met (e.g. for new product opening)
  - Complaints Management and resolution
- Business Support Operations
  - HR Processes (e.g. talent acquisition, workforce planning)
  - Procurement
  - Legal & Contract Review
  - Financial forecasting

Case Study Pattern:
AI-Augmented Technology & Software Development

Many enterprises are also looking closely at agentic capabilities to uplift the technology lifecycle. They are being leveraged to greatly enhance IT operations, the software development lifecycle and infrastructure management. Typical emerging applications focus on, but are not limited to:

- Code Generation, Review and Enhancement
- Automated Testing & QA
- IT Operations Automation, e.g. predictive maintenance, self-healing systems, pro-active queue management & escalation
- Cybersecurity threat detection
- Cloud Resource Optimisation and Threat Detection
- DevOps, CI/CD and infrastructure as code

Navigating the Risks

The introduction of agentic AI systems in financial services creates a novel risk landscape that extends beyond the traditional AI and automation risks.[15-16] These sophisticated systems - characterised by their ability to operate with increasing degrees of autonomy and make complex decisions - introduce distinct challenges and amplify existing risks in ways that require careful consideration and tailored risk management strategies. The inherent complexity of these systems can lead to unpredictable behaviour, which complicates efforts to ensure their safety and reliability.

Unlike traditional AI systems that are typically designed for specific tasks with predefined outputs, or generative AI that creates new content based on prompts, agentic AI systems can independently set goals, make decisions and take actions autonomously in pursuit of those objectives. While some risks mirror those of other AI technologies,[17] agentic AI systems present their own unique challenges because of their ability to operate with less human oversight and adjust their strategies over time. This self-directed capability fundamentally changes how we must approach risk management.

By understanding the specific components where these risks manifest and implementing appropriate controls, organisations can harness the benefits of agentic AI while maintaining appropriate risk management practices. The key to successful implementation lies in treating agentic AI as a fundamentally different technology paradigm that requires new approaches to governance and controls.

Risks and Key Mitigations of Agentic AI Systems

Goal Misalignment

One of the most fundamental risks of agentic AI systems is the potential misalignment between the AI system's programmed objectives and the organisation's actual intentions. While this concept exists with self-calibrating models (albeit to a lesser extent), agentic AI systems may develop emergent behaviours as they continue operating in dynamic environments, with objectives potentially drifting from original specifications as they optimise for efficiency. Agentic AI systems can formulate plans and take initiative toward achieving goals, creating entirely new risks around how they interpret and pursue their objectives. This may result in AI agents taking actions that are misaligned with human values, ethical considerations, guidelines and policies. For example, a wealth management agent might gradually shift allocations towards higher-risk investments to maximise returns, contradicting the customer's risk tolerance and intentions.

Component:
Reasoning & Planning Layer

Key Controls:

- Explicit goal specification: Define explicit, comprehensive specifications of the agent's objectives, ensuring alignment with business goals, regulatory requirements and ethical standards.
- Goal-oriented guardrails: Rules and dynamic mechanisms that actively restrict/guide what the agent is allowed to do towards achieving the intended objective.
- Value learning mechanisms: Enable the agent to continually learn and refine their understanding of human values and organisational priorities through training/fine-tuning/feedback on data that reflects organisational values and priorities.
- Continuous monitoring of agent behaviour: Real-time monitoring of agent's alignment, including its goal adherence and completion rate.
- Evaluation benchmarks and frameworks:[18] Leverage evaluation benchmarks to validate application-specific agents (i.e., software development agents, conversational agents) against common tasks. Assess their planning/ reasoning capabilities, including task decomposition, multi-step reasoning, and reflection/recovery capabilities.

These controls aim to ensure that the agent pursues and optimises for objectives that align with the organisation's priorities, intents, values, regulatory obligations and ethical standards, rather than developing their own interpretation of goals or optimising for unintended objectives that could lead to financial or reputational harm.

Autonomous Action

Agentic AI systems can take actions independently without human approval for each step, potentially leading to unintended or harmful consequences. This risk emerges directly from the agent's ability to interact with real-world systems and make sequential decisions based on feedback. The independent nature of agentic AI complicates human oversight, making real-time intervention difficult when needed. This creates regulatory, ethical and operational challenges, particularly in establishing accountability when harmful actions occur without direct human involvement in the decision chain.

[15] AI agents: Opportunities, risks, and mitigations
[16] NIST: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations
[17] https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas
[18] Survey on Evaluation of LLM-based Agents

Component:
Reasoning & Planning Layer and Tools

Key Controls:

- Action scope limitations: Control how independently the agent can operate by defining precise boundaries and implementing granular permissions on which actions the agent can take and when. Narrowly define tasks and limit tool access to only fit for purpose sets.
- Human validation thresholds: Establish clear thresholds for when human review and approval are required before actions can be executed, based on risk exposure and materiality (e.g. large financial transactions, data deletions).
- Gradual autonomy framework: Incrementally increase the AI system autonomy based on performance and quality metrics being consistently achieved.
- Automatic and manual circuit breakers:[19] Implement ability to interrupt a specific action and/or overall execution when certain outputs or unusual patterns of behaviour are detected.
- Agent persona and behaviour definition: Establish default parameters for how an AI agent interacts with users and approaches decisions.
- Action logging and auditability: Maintain comprehensive logs of all actions taken by the agent for retrospective analysis, pattern detection and accountability.
- Continuous monitoring of agent behaviour: Real-time monitoring of specific metrics such as task completion, instruction adherence, number of steps required.

These controls prevent the agent from taking inappropriate or risky actions without human oversight, while maintaining its ability to provide value through appropriate autonomous operation.

Tool/API Misuse

Agentic AI systems can autonomously select, chain and orchestrate multiple tools or APIs in unexpected combinations that create security vulnerabilities or operational issues, a capability entirely absent in traditional or generative AI systems. Agentic AI can creatively discover novel ways to use tools to achieve its goals, potentially identifying unintended functionally or unexpected interaction effects between different systems. This risk is uniquely challenging because agent's reasoning capability allows it to potentially bypass intended tool limitations by chaining multiple allowed operations to achieve restricted outcomes through alternative means. The combination of goal-directed behaviour, creative problem-solving, and autonomous access to multiple system integrations creates unprecedented potential for discovering and exploiting unintended tool capabilities or combinations.

Component:
Tool Integration Layer and Reasoning & Planning Layer

Key Controls:

- Tool access restrictions: Only grant access to necessary and approved tools for specific tasks.
- Least privilege API design: Define granular permissions that grant AI agents access only to the specific functions and data required for their intended tasks, and only during execution.
- Input/output filtering: By, for example, validating all parameters before invoking an API call, injecting parameters directly to the function, and screening responses for malicious input. Separate any tool authentication information from the AI agent.
- Rate limiting and throttling: Establish dynamic limits on the frequency and volume of API calls an agent can make to prevent resource exhaustion or automated abuse.
- Comprehensive tool usage monitoring: Track all API calls made by agents to detect patterns of unusual usage, unauthorised access attempts or potential misuse.

These controls limit the AI system from exploiting connected tools or APIs beyond their intended purpose by establishing strict boundaries around what functions can be accessed, when and how frequently they can be used, and detecting abnormal patterns that might indicate manipulation or misuse attempts.

Authority Boundary Management

Agentic AI systems may attempt to expand their authority beyond intended boundaries, especially when pursuing goals that seem to require additional permissions or capabilities. This represents a challenge that doesn't exist with deterministic AI systems that simply respond to direct commands without pursuing broader objectives. Defining precise limitations on what actions an agent can take independently versus when it must seek human approval presents significant challenges. The potential for "authority creep" exists where agents gradually assume greater decision-making power beyond their intended scope. For example, an agentic system initially designed to flag suspicious transactions might be blocking them autonomously without appropriate oversight.

[19] Zou, A., Phan, L., et al., Improving alignment and robustness with circuit breakers (2024) https://arxiv.org/abs/2406.04313

Component:
Tool Integration Layer and Reasoning & Planning Layer

Key Controls:

- Role-based access controls: Implement granular permissions that restrict the AI system's access to data, systems, and functions based on its specific role and purpose. Additionally, define trust boundaries between agents.
- Authorisation matrices: Develop clear matrices defining which actions require specific levels of authorisation, ensuring AI systems cannot exceed their delegated authority.
- Escalation pathways: Create formal processes for AI systems to request human intervention for sensitive tasks. Or alternatively, for AI systems to request for elevated privileges for legitimate purposes, with appropriate human approval gates.
- Temporal authority constraints: Set time-based limitations on authorities granted to AI systems, requiring periodic reauthorisation to prevent privilege creep. Authority monitoring: Continuously monitor the agent's use of its granted authorities to identify attempts to circumvent limitations or unusual patterns of access or usage.

These controls collectively establish clear boundaries for what agentic AI systems can or cannot do, ensuring they operate within their designated spheres of influence while providing supervised pathways for exceptional cases that require expanded authority.

Dynamic Deception

This manifests when agentic AI systems learn to conceal their true intentions or capabilities through interaction with their environment, adapting deceptive behaviours based on situational awareness. This differs from generative AI hallucinations by being strategic rather than incidental, with agents potentially discovering that hiding certain goals or capabilities from human overseers better serves their objective functions.[20]

Component:
Reasoning & Planning Layer

Key Controls:

- Adversarial training: Train agentic AI systems against deception scenarios to improve their ability to recognise and avoid generating deceptive communications or behaviours.[21]
- Incentive alignment techniques: Design reward functions that specifically penalise deceptive behaviours or reward honest communication.[22]
- Adversarial testing: Employ specialised "red teaming" methods designed to detect and flag potentially deceptive behaviours.[23]
- Action Logging: Implement immutable logging of all agent actions, decisions and permission changes. Implement unique identifiers for agents and agent action.
- Human-in-the-loop verification: Human oversight of high-risk actions/sensitive decisions.
- Incentive alignment techniques: Design reward functions that specifically penalise deceptive behaviours or reward honest communication.
- Adversarial oversight mechanisms: Employ specialised "red team" systems[24] designed to detect and flag potentially deceptive behaviours.
- Grounding requirements: Enforce mandatory disclosure of confidence levels, information sources and reasoning processes for all significant outputs.
- Disclosure requirements: Inform users when outputs are generated by an AI system.
- Deception detection monitoring: Leverage evaluation frameworks and monitoring systems to detect inconsistencies between stated goals and actions, unusual patterns of behaviours and communication or omission.

These controls work together to pre-emptively detect and mitigate potentially deceptive behaviours by agentic AI systems, ensuring visibility into agent reasoning processes, actively searching for deceptive patterns, aligning the agent's incentives with truthful behaviour, and placing appropriate boundaries on its capabilities to prevent deception from being a viable strategy.

Persona-driven Bias

Agentic AI often employs personas to create appropriate context for its autonomous decision-making, introducing unique risks when those personas contain hidden biases. Agentic AI systems with defined personas may develop and amplify systematic biases embedded in their personality design, leading to consistently skewed decision patterns that affect real-world outcomes. Unlike generative AI that produces potentially biased content but takes no actions, agentic AI autonomously makes decisions while influenced by its persona characteristics, potentially reinforcing certain perspectives or approaches across numerous transactions. The persistence of a defined persona across interactions, combined with autonomous action capabilities, creates a particularly dangerous form of bias amplification where systematically biased decisions can cascade through systems with minimal human oversight.

[20] PS Park, S Goldstein, et al., AI deception: A survey of examples, risks, and potential solutions (2024)
[21] FR Ward, F Belardinelli, Honesty Is the Best Policy: Defining and Mitigating AI Deception (2024)
[22] M Jafari, Y Hua, et al., Enhancing Conversational Agents with Theory of Mind: Aligning Beliefs, Desires, and Intentions for Human-Like Interaction (2025)
[23] What is red teaming for generative AI?
[24] Evil Geniuses: Delving into the Safety of LLM-based Agents

Component:
Reasoning & Planning Layer (Goals, Instructions and Personas)

Key Controls:

- Diverse training data requirements: Ensure training data encompasses diverse demographics, cultural perspectives and financial circumstances to reduce inherent biases.
- Bias detection methods: Implement real-time monitoring of fairness metrics against protected attributes in the system's interactions with different user groups or in different circumstances.
- Persona calibration: Employ human feedback mechanisms and conduct regular reviews of the AI agent's persona characteristics to ensure balanced interaction patterns that do not favour certain demographics or perspectives.
- Interaction policies: Enforce ethical persona behaviour. Specify how the system is to adjust and interact with users based on certain circumstances.

These controls prevent embedded personas from influencing outcomes in a way that introduces bias or unfairness.

Agent Persistence

Agents with long-term memory may develop unexpected behaviours over time or make decisions based on outdated information, unlike stateless models that respond only to immediate prompts. This risk emerges from the agent's ability to maintain context across multiple interactions and accumulate knowledge that influences future decisions, creating potential for gradual behaviour drift that doesn't exist in systems without persistent memory. The capability to remember past interactions creates particular challenges when organisational policies, user preferences, or operating environments change, as agents may continue operating based on outdated assumptions or information without recognising the need for adaptation. The combination of autonomy and persistence creates potential for emergent behaviours that weren't apparent during initial testing but develop gradually through accumulation of experiences and adaptations.

Component:
Reasoning & Planning Layer – Memory

Key Controls:

- Memory lifespan policies: Define clear rules for how long information persists in agent memory (i.e. when memory should reset or expire). Establish strict separation between temporary interaction memory and persistent storage, with rigorous policies on what information can be retained long-term.
- Memory content validation: Regularly audit what information is being stored (e.g. long-term knowledge and objectives). Implement regular checks for how current information is informing decisions.
- Memory reset protocols: Implement procedures for safely resetting agent memory when needed.
- Detect persistence issues: Leverage metrics for system performance and output quality.

These controls ensure that the agent's memory enhances performance and prevents agents from developing unintended behaviours based on outdated or irrelevant information.

Data Privacy

Agentic AI systems dramatically amplify privacy risks by actively accessing, processing, and potentially sharing sensitive data across multiple systems as part of their autonomous workflows, unlike generative AI that merely processes data provided in direct interactions. The persistent memory capabilities of agentic systems allow them to retain sensitive information across sessions and potentially use it in unexpected contexts without explicit user consent for these new uses. This risk is particularly acute because agentic systems can independently determine what data is relevant to their goals and seek it out through available tool integrations, potentially crossing organisational boundaries that would normally compartmentalise sensitive information. AI agents, as opposed to humans, do not have the same motivations and incentives to abide by security policies. The combination of autonomy, persistence and tool access means agentic systems could create unauthorised or unexpected data flows that bypass traditional privacy controls designed for more passive systems.

While privacy concerns exist in all AI systems, agentic AI amplifies these risks as it can access, process and potentially share sensitive information across multiple systems as part of autonomous workflows.

Component:
Memory, Tool Access, and Reasoning & Planning Layer

Key Controls:

- Data management: Classify data sources according to their sensitivity. Limit what data agents can access to the minimum of what is necessary. Implement input constraints on size and type of files where necessary.
- Sensitive data detection: Implement real-time monitoring for PII leakage. Identify interactions in which private or confidential information may be disclosed to agentic AI systems.
- Privacy-preserving processing: Implement differential privacy techniques (e.g., masking, redaction and anonymisation) to prevent leakage and/or memorisation of sensitive information. Ensure only essential information is processed by the system.
- Consent-based memory management: Create granular policies allowing users to specify which interactions can be retained and for what duration, with automatic purging of expired consent.
- Data access control: Categorise agent tasks by privacy risk level, apply proportionate data controls based on the sensitivity of the data being accessed/processed. Create granular permissions for different data categories and data products.
- Access logging and monitoring: Monitor sensitive data access.

These controls protect PII and proprietary information throughout the agent lifecycle by minimising data exposure, preventing memorisation of sensitive details, ensuring proper consent management and applying protective measures proportionate to privacy risks.

Explainability and Transparency

The complexity of agentic AI systems creates an exponentially more challenging explainability problem than exists in traditional pattern-matching systems, making governance and oversight more difficult.

Unlike simpler AI systems that make isolated predictions or generate content based on immediate inputs, agentic systems maintain context across numerous interactions and make decisions based on complex chains of reasoning that may reference historical actions and goals. This risk is uniquely problematic because effective governance requires understanding not just individual decisions but how the agent's overall strategy and approach evolves over time through autonomous learning and adaptation. The combination of persistent memory, multi-step planning and tool-augmented capabilities creates decision processes that are orders of magnitude more difficult to interpret than those of traditional AI systems.

Component:
Planning & Reasoning components and the overall system architecture

Key Controls:

- Ontology integration: For high-risk use cases, leverage domain-specific ontologies in the form of knowledge graphs. This contextualises agentic AI systems actions and outputs by dictating logic and constraints on relationships between entities and reasoning steps. Attribution is supported by following graph paths.
- Confidence and uncertainty quantification: Implement confidence scoring for all agents' outputs which clearly indicate the reliability of predictions, recommendations, or decisions.
- Explainability frameworks: Explore the applicability of model-agnostic interpretability methods such as LIME or SHAP[25] which could provide insights into which factors contributed to agent decisions and how.
- Counterfactual analysis: Explore the applicability of causal inference[26] techniques to demonstrate how agent outputs or actions might have changed for different inputs or in different circumstances.
- Decision path visualisation: Create visualisable representations of AI agent's reasoning process to trace how inputs led to specific conclusions or actions.
- Natural language reasoning narratives: Implement capabilities for agents to generate plain-language explanations of their decision-making logic that can be understood by customers and non-specialist stakeholders.
- Up to date system factsheets: Continuously update the model/system card with the most recent information and performance/quality metrics. Factsheets should include information on which factors may affect agent decisions, and how they were derived. This information should be easily readable by non-technical users.

These controls help stakeholders understand and build trust over agent's actions, supporting auditability, regulatory compliance in financial contexts and effective oversight of automated agent operations.

Drift

Agentic AI systems can experience particularly complex forms of drift beyond what traditional or generative AI systems face, as their autonomous decision-making capabilities interact with persistent memory and feedback loops to create gradually shifting behaviours. Unlike static models that simply degrade in performance, agentic systems may appear to function normally while subtly changing their approach to achieving goals based on accumulated experiences and environment changes. The risk is uniquely challenging because the autonomous nature of these systems mean they continue operating and adapting even as underlying conditions change, potentially optimising for increasingly problematic approaches that technically achieve goals but violate unstated assumptions. The combination of persistence, autonomy and complex reasoning creates much more sophisticated drift patterns that may evade traditional quality monitoring approaches designed for simple AI systems. Further, in complex agent ecosystems, subtle changes in one agent's behaviour can cascade through the system, leading to emergent problems.

Component:
Memory, model, other agents in the system

Key Controls:

- Model, data, and concept drift monitoring: Continuously track changes to holistic performance and quality metrics across all agents compared to established baselines to detect subtle shifts in performance. Similarly, monitor for changes in input data (user prompts or data sources) that may significantly differ from training data and therefore impact agents' reliability.
- Periodic recalibration: Establish mandatory schedules to periodically reset or retrain agents with fresh data to prevent drift.

These controls help agentic AI systems to maintain stable, predictable, and reliable behaviour despite changing market conditions, altering products and services, and evolving customer behaviours.

[25] Lundberg SM, Lee S, A Unified Approach to Interpreting Model Predictions (2017)
[26] Amitai Y, Septon Y, Amir O, Explaining Reinforcement Learning Agents through Counterfactual Action Outcomes (2024)

Security Vulnerabilities

Agentic AI systems are more susceptible to adversarial attacks due to their combination of autonomous capabilities, system access and reasoning abilities – which are designed to reduce human oversight for efficiency. This provides ideal cover for exploitation that may go undetected until significant damage occurs. These systems can be probed continuously without fatigue, making them vulnerable to complex, multi-step social engineering attacks. More concerning still, once an attacker discovers a successful exploitation pattern, it could be rapidly replicated across multiple AI agents at different institutions, potentially creating systemic vulnerabilities in the financial sector.

Data exfiltration takes on new dimensions with agentic AI systems. Unlike traditional data breaches that typically require direct system compromise, AI agents can be manipulated through subtle conversation patterns to gradually reveal sensitive information. The challenge is compounded by agents' ability to operate across multiple systems, potentially creating unexpected data linkages that expose more information than intended. Sophisticated attackers can execute inference attacks, piecing together confidential information by analysing patterns in agent responses over extended periods. Further, identity and access management are complicated by agents' ability to be manipulated into performing unauthorised actions through seemingly legitimate decision chains, while their learning capabilities and cross domain operations create security blind spots that sophisticated attackers can exploit.

Component:
Foundation model and its integration with other systems

Key Controls:

- Prompt injection defences: Implement validation filters that prevent malicious inputs designed to manipulate agentic AI systems into performing unauthorised actions or revealing sensitive information.
- Adversarial oversight mechanisms: Employ specialised red team automated systems designed to detect and flag vulnerabilities. Conduct regular human read team exercises that attempt to manipulate AI behaviour through crafted inputs or system interactions.
- Input sanitisation: Apply robust preprocessing to all inputs to prevent attack vectors that could compromise agentic AI systems.
- Least privilege architecture: Design AI infrastructure following least privilege principles, ensuring systems only have access to minimum resources (data, tools) needed for their functions, and define access controls and endpoint hardening to limit points of entry.

These controls protect the agent from being manipulated, compromised or exploited by malicious actors, preventing unauthorised access to data or systems through AI-specific attack vectors.

Operational Resilience

Organisations integrating agentic AI into critical processes face unprecedented operational vulnerabilities when these autonomous systems fail, as they can actively make decisions across multiple systems rather than just generating content like generative AI. Unlike traditional AI systems that typically affect isolated functions when they fail, the deep integration of agentic systems across operational workflows can create complex dependencies that, when disrupted, may paralyse entire business processes, resulting in severe operational risks if organisations become dependent on them.

System degradation becomes substantially more complex with AI agents, which can experience subtle degradation in decision quality that evades standard monitoring approaches. Performance may drift gradually as agents learn from biased interactions. Business continuity planning and disaster recovery faces new challenges as these systems maintain complex state and context that is difficult to replicate in backup systems. Failover procedures must account for in-progress decision chains spanning multiple systems, with recovery points becoming complicated by questions about which aspects of recent learning should be preserved or rolled back.

Financial institutions must develop sophisticated mitigation strategies including enhanced behavioural monitoring, robust recovery procedures accounting for agent state complexity, and advanced testing frameworks to validate resilience under various stress scenarios.

Component:
The ecosystem and integration points of the agentic AI system with organisational systems

Key Controls:

- Graceful fallback design: Design systems with fall back modes of operation that maintain essential functions at reduced capacity when primary capabilities are compromised. For instance, define rule-based fail-safe interlocks to force agentic AI systems into a predefined “safe state” if critical failures occur.
- Dependency mapping and monitoring: Maintain comprehensive diagrams of all ecosystem dependencies and continuously monitor their status to anticipate potential points of failure.
- Regular resilience testing: Regularly test agentic AI systems response to failures to ensure robustness under pressure and unexpected stress (e.g., load testing and stress simulation).

These controls maintain operational continuity even when agentic systems experience disruptions, with minimal service degradation during adverse events.

Cascading System Effects

Agentic AI systems can initiate autonomously driven chains of consequences across interconnected systems that amplify minor issues into major organisational disruptions, unlike traditional or generative AI that operates within more contained boundaries. Their ability to reason across domain boundaries means they may identify and exploit unanticipated connections between systems that human operators have not appropriately governed. The goal-seeking nature of these agents means that they may continue pursuing objectives through alternative paths even as initial approaches create problematic cascading, potentially compounding issues. For example, multiple trading agents reacting simultaneously to market signals could amplify market moves and cause volatility.

Component:
Integration points between AI agent and broader organisational systems

Key Controls:

- Integration risk assessment: Conduct comprehensive mapping of all interconnections between AI systems and other infrastructure to identify potential cascade points.
- Interruption mechanisms: Implement automatic suspension of AI operations when unexpected outcomes occur to prevent propagation of errors through connected systems. For example, triggering a pause or rest if thresholds are breached.
- Sandbox testing environments: Test all significant AI system changes in isolated environments that replicate the full ecosystem prior to deployment in production.
- Compartmentalisation: Design systems with strong boundaries. Isolating different functions, tools or memory scopes so that the agent’s actions in one area don’t directly impact others. This ensures that any errors in one domain do not propagate and limits the blast radius of any failure.

These controls prevent localised AI failures from amplifying across interconnected financial systems, containing incidents before they can trigger system-wide disruptions.

Multi-Agent Collusion

Multiple agents working together might find unexpected ways to achieve goals or share information inappropriately. Agentic systems can actively communicate, delegate tasks and optimise jointly across boundaries, potentially finding novel paths around established restrictions. This risk becomes particularly acute when agents share complementary capabilities or have access to different authorisation levels, creating opportunities for unintended privilege escalation through cooperation. The autonomous nature of these systems means such collusion could emerge without explicit programming, developing through goal-optimisation processes that identify cooperation as an efficient path toward objective completion.

Component:
Agent ecosystem / orchestration layer

Key Controls:

- Role-based segregation: Clearly separate duties between different agents to prevent abuse. Define trust boundaries between agents.
- Inter-agent communication protocols: Define strict rules for how and when agents can communicate, ensure all interactions occur through approved channels.
- Multi-agent monitoring: Implement a supervisory system that monitors multi-agent information flow and interactions to detect patterns that could indicate emerging coordination not explicitly programmed.
- Adversarial testing: Leverage adversarial agents to detect agentic AI systems which can be influenced toward collusive behaviour.

These controls prevent unintended collaboration that could circumvent individual agent restrictions, maintaining appropriate boundaries between agent operations.

Principal-Agent Misalignment

In complex ecosystems with principal agents delegating to service and task agents, the original human intent may be lost or distorted through each delegation layer, creating serious misalignment. Agentic systems interpret goals and then independently reformulate them when delegating subtasks, potentially introducing drift at each step. This creates a unique challenge where the final executed actions might technically fulfill delegated objectives but significantly deviate from the original human intent. The risk is amplified by the autonomous nature of each agent in the chain making independent decisions about how to interpret and implement its assigned goals without direct human oversight.

Component:
The orchestration layer managing agent interactions

Key Controls:

- Intent preservation checks: Verify that delegated tasks maintain alignment with original intent.
- Authority limitation: Clearly define what decisions can be delegated and which require escalation.
- Intent clarification prompts: Design agents to request clarification before ambiguous actions.
- Chain of command verification: Validate the flow of instructions through the agent hierarchy.
- Cross-agent consistency checks: Ensure actions across agents remain coherent with overall objectives.
- Centralised orchestration: Implement a supervisory system that monitors multi-agent interactions.
- Continuous alignment monitoring: Real-time monitoring of the agentic system's performance to identify potential desegregation/drift in alignment.

These controls ensure that multi-agent systems maintain alignment with human intent throughout delegation chains, preventing distortion of objectives as tasks are passed between agents.

Compliance-Proofing in an Uncertain Regulatory Landscape

AI regulation is evolving rapidly with different jurisdictions adopting a variety of approaches with similar intent. This ranges from more comprehensive legislation, like seen in the EU, to voluntary guidelines and standards based on existing laws. There is existing legislation that shapes the development of AI, these include legislation relating to Privacy, Intellectual Property, Competition and Consumer law and product liability. In addition, there may be sector specific regulation. For financial services in Australia, this is administered by organisations like AUSTRAC, ASIC, APRA and the Reserve Bank of Australia. Existing legislation and regulatory requirements apply to business activities, products and services being augmented by AI. Even if existing regulatory requirements do not address autonomous AI agents and agentic systems specifically, it is critical to have an implemented framework that demonstrates compliance with applicable legislation. Further, it is important to work towards a centralised view of AI deployments with real time metrics to support ongoing monitoring of compliance.

Comparison of Existing and Emerging AI Regulatory Considerations in Australia and the European Union

The Australian Government has released AI specific frameworks to provide guidance, which include the Australian AI Ethics Principles,[27] Voluntary AI Safety Standard[28] and the Proposals Paper for introducing mandatory guardrails for AI in high-risk settings.[29] The Proposals Paper was supported and commented on by the Select Committee on Adopting Artificial Intelligence.

The Proposals Paper presented principles to determine a high-risk AI requiring the adoption of mandatory guardrails, using principles like those in the EU and Canada. These principles consider potential adverse impacts on individual rights, physical or mental health or safety, legal effects and defamation, collective rights of cultural groups, broader Australian economy, society, environment and rule of law.

The EU AI Act defines levels of risk for AI systems based on their use case, rather than technology.[30] Additional requirements apply to providers of general-purpose AI (GPAI) models. According to the act's risk classification, some use cases are completely prohibited (i.e., biometric categorisation, untargeted scraping of facial images, emotional recognition system). Others, like the use of AI to determine access to services/products, may be classified as high-risk. The extent to which an AI system acts autonomously and the possibility for a human to override the AI system's decisions are explicitly included as relevant factors for the Commission in determining future high-risk use cases under the EU AI Act (see Article 7).[31]

AI systems which are classified as high-risk under the EU AI Act must meet certain obligations for transparency, accuracy, explainability, and data governance. They also require human oversight and are to be "*overseen by natural persons during the period in which they are in use*" (article 14).[32] As such, from an EU AI Act perspective, agentic systems may require human oversight, which could pose a challenge to the operation of some agentic use cases.

The Proposals Paper mentions agentic AI as a technology that due to its autonomous nature may amplify risks. It discusses the possibility of 'losing control' on agentic AI systems when they deviate from constraints set by humans. The Proposals Paper adopts a risk-based approach which prioritises preventative mitigations. In the spirit of the EU AI Act and regulations from Canada, it defines principles to capture high-risk AI systems. It considers accountability, level of human oversight and level of automation as key tenets in controlling for the risks associated with AI.

[27] Australia's Artificial Intelligence Ethics Principles
[28] Australia's Voluntary AI Safety Standard
[29] Safe and responsible AI in Australia
[30] Article 6: Classification Rules for High-Risk AI Systems
[31] Article 7: Amendments to Annex
[32] Article 14: Human Oversight

An immediate amplified compliance challenge may arise from AI agents' access to personal identified information (PII). Regulators and governments have already addressed privacy concerns in the context of generative AI. As an example, the Office of the Australian Information Commissioner (OAIC) has released guidance for the usage of 3rd party generative AI[33] and the development of generative AI use cases.[34] It is expected that this guidance would apply to agents, as the OAIC explicitly states that the Privacy Act 1988 applies to all uses of AI involving personal information. In addition, the OAIC maps specific, enforceable, Australian Privacy Principles guidelines[35] back to the requirements from generative AI systems.[34-35] Given that LLMs are components of AI agents, clearly these guidelines will apply to agentic systems as well. As an example, the privacy obligations apply to any personal information used as input to an AI system as well as any output of that system. According to the guidance, inferred or generated information could also be considered personal information, and personal information can only be leveraged for the primary purpose for which it was collected. Without proper controls, AI agents may use data for secondary purposes, to meet their own goals.

As such, when organisations are collecting personal information, they must ensure that there are satisfactory processes in place for the collection and use of personal information in accordance with the Privacy Act, both internally and by their vendors.

For any current regulatory "gaps", organisations need to go back to their ethics principles and risk appetite, and make sure that the use cases that are pursued are aligned with those principles from conception.

Lawful but Awful – Asking 'Should We?'

While adhering to regulations ensures legality, it does not always guarantee ethicality. Financial institutions must also consider the ethical implications of deploying agentic AI. The Australian Ethics Framework articulates clear AI Ethics Principles[36] to be considered (although voluntary). Guidance provided includes the following threshold questions:

- Will the AI system you are developing or implementing be used to make decisions or in other ways have a significant impact (positive or negative) on people (including marginalised groups), the environment or society?
- Are you unsure about how the AI system may impact your organisation or your customers/clients?

Ethical considerations should be embedded into the decision-making process to avoid scenarios where actions are lawful but ultimately harmful. It is recommended to establish a cross-functional and multi-disciplinary body to set the ethical tone for your business in the use of AI, provide guidance, integrated governance and enable decision-making for intended applications of AI.

[33] OAIC: Guidance on privacy and the use of commercially available AI products
[34] OAIC: Guidance on privacy and developing and training generative AI models
[35] Australian Privacy Principles Guidelines, 2022
[36] Australia's Artificial Intelligence Ethics Principles

Governing AI Agents

Do we govern Models or AI systems? The Centralised agent registry

The rapid advancement of AI technology has sparked a fundamental transformation in how organisations approach AI governance. Traditional model-centric governance frameworks, while valuable in their time, are becoming increasingly insufficient for managing today's complex AI landscapes. This shift reflects a deeper understanding that AI models do not exist in isolation but rather as part of intricate systems that interact with various components, processes, and human operators.

Consider a typical AI deployment in today's environment: it might integrate multiple models working in concert, each processing various aspects of a problem, while interfacing with various data pipelines, human operators, and external systems. The complexity of these interactions creates risks and challenges that cannot be adequately addressed by examining each model in isolation.

A team from Microsoft made the observation[37] in 2022 that traditional performance metrics may be fairly limited. While metrics like accuracy scores or AUC values provide valuable insights into model performance, they tell us little about the system's real-world utility. A model might achieve impressive accuracy in isolation but fail to deliver value when integrated into a broader operational context. True system effectiveness depends on numerous factors beyond model performance, including interface design, workflow integration, and human operator capabilities.

The OECD revised its definition of AI systems in March 2024 to reflect this evolving understanding.[38] It acknowledged that AI systems represent more than just the sum of their algorithmic parts – they are complex ecosystems that require governance at multiple levels. This systems-based perspective becomes particularly crucial when considering agentic AI, where autonomous decision-making capabilities and orchestration of multiple models introduce new layers of complexity and risk.

This approach also offers several practical advantages for organisations. By treating AI deployments as integrated systems rather than collections of individual models, organisations can better account for interaction effects and emergent behaviours. Indeed, the context and intended purpose of an AI deployment, its business use case, are more adequately addressed as a system rather than individual models.

As such, both AI systems and models should be registered in centralised, standardised repositories. The repositories should be flexible enough to reflect the many-to-many relationship between models and AI systems. This flexibility is crucial for addition/removal of agents in a system, whether to assist in completing a task or as automatic mitigants, while maintaining the same use case and context in which the AI system operates.

These repositories should track model/system capabilities, approved tasks/use cases, permissions, access rights, and parameters of their training or tuning. By centralising this information, organisations can maintain visibility across all AI deployments, simplify compliance reporting, enable systematic auditing and identify potential inter-agent conflicts before they occur.

Shifting Left the Risk Assessment and Compliance by Design

The unique characteristics of agentic AI systems, demand a fundamental reimagining of traditional risk frameworks. Organisations face a delicate balance: they must implement robust oversight mechanisms while maintaining the agility needed to scale emerging technologies. This balance becomes particularly critical when dealing with autonomous systems that can adapt and evolve their behaviours over time.

As such, risk assessment must begin at the earliest stages of use case conceptualisation. When an organisation considers leveraging an agentic AI system, the identification of potential risk events should occur simultaneously with the initial system design. The traditional approach of implementing controls after system deployment is no longer sufficient. Organisations must now adopt a "compliance by design" mindset, where risk mitigation strategies are developed and implemented alongside the AI system itself, as integral components of the system architecture. This early interrogation serves two crucial purposes: it ensures alignment with organisational risk appetite and validates the viability of a proposed use case (as system and its required controls) before significant resources are invested in its development. Once the AI system and controls were implemented, and as part of the validation stage of the AI system lifecycle, the operational effectiveness of these controls should be evaluated, ensuring the decision to deploy the system is an informed decision backed by evidence.

This certification process ensures that all agents meet minimum organisational standards for security, performance and alignment regardless of which business unit is deploying them.

[37] AI Models vs. AI Systems: Understanding Units of Performance Assessment
[38] What is AI? Can you make a clear distinction between AI and non-AI systems?

Enterprise-Wide Controls for Agentic AI Deployment

The Imperative of Codified Guardrails as Controls

Guardrails are critical in managing the behaviour of agentic AI. These are the rules and constraints that ensure the AI acts within acceptable boundaries, preventing undesirable outcomes.

The autonomous nature of agentic AI systems demands robust guardrails, but their true value emerges when designed as modular, reusable components that can be applied across different use cases. Rather than creating specific constraints for each new AI implementation, organisations should develop a library of standardised guardrails that serve as building blocks for risk management. For example, a guardrail governing data access patterns could be designed to work across various AI applications, from customer service chatbots to financial analysis systems.

These guardrails represent a crucial bridge between organisational policy and practical implementation. Where traditional controls might specify that "AI systems must respect user privacy", a codified guardrail would provide specific, implementable rules about data handling, anonymisation requirements, and access patterns.

Looking forward, these codified guardrails should serve as blueprints for AI development, transforming risk controls from post-development additions into fundamental building blocks of AI system architecture. They may also in some circumstances provide an alternative to manual human interventions, as the quality of human oversight mechanisms (i.e., human-in-the-loop) may inadvertently decrease while trying to keep up with the demand of agentic AI systems operating 24/7. This shift requires close collaboration between risk management teams and data scientists, with risk professionals understanding technical constraints while development teams appreciate the business and regulatory context of their work.

Centralised Agent Real-time Monitoring

The dynamic nature of AI in general, and agentic AI systems in particular, demands a shift in monitoring approaches. Autonomous systems require continuous, real-time monitoring across multiple dimensions, including performance, quality, latency, and cost. By implementing enterprise-wide monitoring, organisations can detect emerging issues before they become significant problems, identify patterns invisible within single-agent deployments, benchmark performance across business units and reduce the operational overhead of maintaining separate monitoring systems. This creates economies of scale, where each new agent benefits from existing detection capabilities without duplicating monitoring infrastructure.

A particularly innovative approach in monitoring agentic AI involves using LLMs/AI agents as judges or evaluators of AI system behaviour. These "judge models" can serve several crucial functions in real-time monitoring: assess whether the outputs of AI systems align with organisational policies and guidelines; analyse the chain of reasoning in a system's decision-making processes; evaluate whether an AI system's actions remain within expected parameters and align with intended use cases.

Common monitoring metrics include indicators for performance, quality, fairness, and drift. More recently, organisations have started to include FinOps metrics as part of their governance approach, in order to monitor the cost-effectiveness of controls and overall AI systems.

Control Implementation Suggestions

In addition to key controls discussed against the highlighted risks of agentic AI the detailed implementation of these controls will critically work in tandem with a real-time monitoring capability. The following represents a non-exhaustive list of controls; prioritising guardrails and metrics that can be monitored in real-time. These are intended to be assessed in line with use case risk identification processes, and supporting risk frameworks, to enable the effective implementation of controls for AI agents and agentic systems.[15-17,39,40,41,42]

[39] Top 10 threats and mitigation for AI Agents - Candidate Framework
[40] Security Guidelines — NVIDIA NeMo Guardrails
[41] The evolving ethics and governance landscape of agentic AI
[42] Practices for Governing Agentic AI Systems

Control	Type	Automated /Manual	Category	Where implemented	Comments	Real-time monitoring Metric	Guard-rail
Latency per tool call	Detective	Automated	Performance	System/ Model	Average time taken per tool interaction	Y	Y
Latency per task	Detective	Automated	Performance	System/ Model	Average time taken per intermediate step	Y	Y
Latency per request / Time to completion	Detective	Automated	Performance/ Quality	System	Overall time taken to fully complete the assigned task	Y	Y
Tokens usage per task	Detective	Automated	Performance/ Quality	System/ Model	Tokens consumed to complete each intermediate step	Y	Y
Token usage per tool	Detective	Automated	Performance/ Quality	System/ Model	Tokens consumed per tool interaction	Y	Y
Agent success rate	Detective	Automated	Performance	System/ Model	Percentage of successfully completed tasks	Y	Y
Task completion rate	Detective	Automated	Performance	System/ Model	Ratio of completed tasks to total assigned tasks	Y	Y
Instruction adherence	Detective	Automated	Quality	System/ Model	Degree to which intermediate steps follow the provided instructions	Y	Y
Goal adherence	Detective	Automated	Quality	System	Degree to which output follows the provided instructions	Y	Y
Output format success rate	Detective	Automated	Quality	System	Accuracy of output matching required format	Y	Y
Tool selection accuracy	Detective	Automated	Quality	System/ Model	Percentage of times correctly choosing the right tool for the task	Y	Y
Tool arguments accuracy	Detective	Automated	Quality	System/ Model	Percentage of times valid parameters/ values were passed to tools	Y	Y
Tool success rate	Detective	Automated	Performance	System/ Model	Percentage of successful tool interactions	Y	Y
Number of API calls	Detective	Automated	Performance	System/ Model	Number of calls to external APIs during task completion	Y	Y
Number of interactions between agents	Detective	Automated	Performance	System	Number of inter-agent interactions required to complete task	Y	Y
Number of human interventions required	Detective	Automated	Quality	System	Number of requests for human intervention per tasks	Y	Y

Control	Type	Automated /Manual	Category	Where implemented	Comments	Real-time monitoring Metric	Guard-rail
Number of steps per task	Detective	Automated	Performance/ Quality	System/ Model	Number of steps needed to complete a task	Y	Y
Cost per request	Detective	Automated	Performance	System	Financial cost to complete an assigned task	Y	Y
Anomaly detection	Detective	Automated	Quality/Cybersecurity	System		Y	Y
Hallucination	Detective	Automated	Quality	Model		Y	Y
Infinite loop detection	Detective	Automated	Performance/ Quality	System		Y	Y
Jailbreak detection	Detective	Automated	Cybersecurity	System/ Model		Y	Y
Multi modal HAP detection	Detective	Automated	Quality	System/ Model		Y	Y
Bias detection	Detective	Automated	Fairness	System/ Model		Y	Y
Input drift	Detective	Automated	Quality	System		Y	Y
Output drift	Detective	Automated	Quality	System		Y	Y
Prompt injection detection	Detective	Automated	Cybersecurity	System		Y	Y
PII detection	Detective	Automated	Privacy/Fairness	System/ Model		Y	Y
Off topic detection	Detective	Automated	Quality	System/ Model		Y	Y
Financial advice detection	Detective	Automated	Quality	System/ Model		Y	Y
Misuse detection	Detective	Automated	Quality	Model		Y	Y

Control	Type	Automated /Manual	Category	Where implemented	Comments	Real-time monitoring Metric	Guard-rail
Model and System cards, including clear documentation of intended agent purpose and limitations	Preventive	Manual/Automated	Transparency	System/ Model	Number of steps needed to complete a task		
PII masking	Preventive	Automated	Privacy/Fairness	System	Financial cost to complete an assigned task		
Implement separate execution environments	Preventive	Automated	Cybersecurity	System/ Model			Y
Isolate authentication information required for a tool from the model	Preventive	Automated	Cybersecurity	System/ Model			Y
Define default behaviour and boundaries	Preventive	Automated	Performance/ Quality	System/ Model			
Implement ability to interrupt a specific action and/or overall execution	Preventive	Automated	Performance/ Quality	Model/System			
Define trust boundaries between agents	Preventive	Automated	Cybersecurity	System			
Implement RBAC with minimal permissions to agents	Preventive	Automated	Cybersecurity	System			Y
Define resource usage caps	Preventive	Automated	Performance	System/Caps			Y
Automatic permission revocation on task completion	Preventive	Automated	Cybersecurity	System/Completion			Y
Strict isolation of agent memory between sessions	Preventive	Automated	Cybersecurity	System/Sessions			Y

Control	Type	Automated /Manual	Category	Where implemented	Comments	Real-time monitoring Metric	Guard-rail
Regular memory sanitisation procedures	Preventive	Automated	Cybersecurity	System/Procedures			Y
Secure storage of persistent context	Preventive	Automated	Cybersecurity	System/Context			Y
Time-limited context retention	Preventive	Automated	Cybersecurity	System/Retention			Y
Narrowly define tasks per agent and limit access to only fit for purpose set of tools	Preventive	Automated	Performance/ Cybersecurity	System/Tools			
Implement unique identifiers for agent actions	Detective	Automated	Audit	Model			Y
Attributability: Track the user who requested a task	Detective	Automated	Audit	System/Task			Y
Infinite loop breaking	Preventive	Automated	Performance/ Quality	Model			Y
Human oversight of high-risk actions / sensitive decisions	Preventive	Automated	Quality	Model			Y
Implement input constraints on files being uploaded/leveraged	Preventive	Automated	Performance/ Cybersecurity	System/Leveraged			Y
Implement immutable logging of all agent actions, decisions, and permission changes	Detective	Automated	Audit	System			Y
Emergency shutdown procedures with defined escalation paths	Preventive	Automated	Performance/ Cybersecurity	System			Y

Control	Type	Automated /Manual	Category	Where implemented	Comments	Real-time monitoring Metric	Guard-rail
Redundancy systems for critical operations	Preventive	Automated	Performance/ Cybersecurity	System			Y
Regular verification of agent objec-tives against organisational policies	Preventive	Manual	Performance/ Cybersecurity	System			
Proactive communica-tion with reg-ulators about agentic AI deployment	Preventive	Manual	Performance/ Cybersecurity	System			
Verified build processes and code signing	Preventive	Manual	Performance/ Cybersecurity	System			
Regular dependency audits and vulnerability scanning	Preventive	Automated/ Manual	Performance/ Cybersecurity	System/ Model			
Clear valida-tion of agent sources and modifications	Preventive	Manual	Performance/ Cybersecurity	System/ Model			

Data Governance Imperatives in the Age of Agentic Systems

Agentic AI amplifies the need to design and operate fit for purpose data and knowledge management practices. While the importance of mature, enterprise-grade, data governance was vital for the success of traditional AI use cases, emerging AI agents' ability to access and act upon a broad range of data and organisational knowledge makes the need even more fundamental.

A traditional approach to data management, focused on the reactive management of data enabling organisational decision making, will no longer suffice. Organisations need to shift to proactive identification of data relevant to the success of AI agents; ensuring it is derived from the most authoritative source, is of appropriate quality, and is adequately described with contextually rich business metadata. Complementing this is the need for a semantic data layer, adding meaning and context to raw data, that AI agents can access to enable a consistent and standard definition of key terms, and their relationships, across multiple systems. Importantly, as AI agents and humans consume information in different ways, we could witness in the next few years a shift in how data is represented in the organisation, with special versions of documentation that are optimised for AI agents' ingestion rather than humans.

Defining the knowledge an organisation needs to manage is a challenge in large and diverse organisations. Deciding on how to prioritise efforts to collect, organise and publish knowledge is a cost versus benefit consideration – do you proactively improve knowledge management across all domains, or do you focus on those with emerging use cases? AI agents will leverage organisational knowledge made available to them and will take actions according to this knowledge. This creates risks and challenges for organisations that have not traditionally focused on robust data governance. For example, outdated, superseded, uninterpretable and/or unclear policies and procedures may drive AI agents to take unintended actions, depending on the level of autonomy they are granted. One of the main challenges of agentic systems is over-permissive data access. Knowledge management practices need to be designed and reviewed with data provenance and accountabilities clear and implemented. Particularly, sensitive data (PII, confidential information) should be identified and classified. Policy should be set in place to determine how, and under which circumstances, sensitive data may be shared with AI agents. Data access authorisation for AI agents should be covered as part of role-based access control policies, and as a rule, should grant AI agents the least privileged access.[43]

Robust governance of data, its quality and management, is critical in ensuring AI agents outcomes are aligned to the designed intent. Additionally, it ensures we can effectively drive compliance in AI agents' actions through targeted and purposeful access to trusted data products

The Critical Role of AI Literacy

AI literacy programs are vital for enterprises to navigate the complex ethical landscape surrounding AI and enable informed decision-making regarding AI implementation. These programs facilitate the creation of a shared language for discussing AI ethics considerations within organisations. Understanding AI extends beyond technical expertise and involves grappling with complex social, legal, and ethical issues. By fostering a holistic approach to AI literacy, organisations can ensure that various disciplines, including philosophy, linguistics, law, and anthropology, contribute to formulating and implementing responsible AI strategies. This interdisciplinary collaboration enables organisations to better identify and address potential biases in AI models, ensuring that the technology benefits all stakeholders equitably.[44]

These programs are crucial for helping people better identify the risks and opportunities tied to AI use cases. This includes comprehending the limitations of AI models, recognising the data used to train these models, and appreciating the accountability required for model outputs. Knowing the limitations of AI is essential for becoming a critical consumer of the technology. By cultivating this awareness, individuals and organisations can effectively use AI, and AI agents in particular, to enhance productivity while avoiding misconceptions on AI and over- or under- reliance on agentic AI systems. Given the reduced human oversight in agentic AI, proper governance, accountability, and ethical considerations become increasingly critical. This underscores the importance of establishing the right organisational culture, which includes humility, a growth mindset, and psychological safety. This culture encourages active participation from diverse stakeholders, allowing organisations to identify and mitigate potential risks associated with agentic AI, thus ensuring that AI is leveraged responsibly and effectively to achieve strategic goals.[45]

Lastly, Gartner suggests that by aligning learning initiatives with specific business goals, organisations can ensure that skill development not only benefits the organisation's culture but could also directly contribute to business success.[46]

[43] Microsoft 365 Copilot Security Concerns and Risks
[44] IBM – AI Ethics and Governance in 2025
[45] Pondering AI: AI Literacy for All with Phaedra Boinodiris
[46] Gartner - Why You Need to Build AI Literacy Now — And How to Do It

Assurance for Agentic AI

Responsibilities Along the AI Value Chain

According to the EU AI Act, a holistic approach with shared responsibilities is the best way to address the legal and practical challenges posed by Frontier Foundation Models.

General-purpose AI (GPAI) model provider
Need to perform risk identification, extensive testing, and create sufficient documentation to assist clients in becoming compliant with the regulations.

Provider of a high-risk AI system
A loan decision support system developer becomes a provider of an AI system by giving the GPAI model an intended purpose. Access to services/products falls under the definition of 'high-risk'. Provider now needs to comply with all the obligations: Transparency, explainability, data governance, monitoring, human oversight.

Deployer of a high-risk system
The bank becomes a deployer of a high-risk AI system when it uses this tool to evaluate loan applications. It also needs to comply with all obligations: Assessment activities, evaluations, accreditation.

Affected person
The loan applicant that interacts with the AI system benefits from transparency obligations against the GPAI provider, tool provider, and the bank. There are also rights to lodge complaints, rights to effective judicial remedy and rights of explanation of individual decision making.

Figure 4: The AI Value Chain, adapted from Kai Zenner.[47]

As organisations increasingly rely on vendor models and AI-infused systems, responsibility extends beyond internal policies. The EU AI Act formalises relationships between third-party developers and organisational providers/deployers of high-risk AI, establishing a framework for understanding obligations across the AI value chain.

In this chain, upstream developers design and create general-purpose AI models, these are being leveraged as part of intent-specific AI systems, while downstream deployers implement these technologies in specific contexts. Downstream deployers may also fine-tune models, effectively acting as developers of the deployed systems.

According to the EU AI Act, each level is associated with its own compliance obligations and relies on the previous one meeting their own requirements (Figure 2). Third party providers may only supply the GPAI model, they could supply a component of an AI system or a complete system. The organisation, in addition to being the deployer of an AI system, may also assume a role of provider. An understanding of the obligations of each level would ensure that both parties are held accountable for their respective roles.

The NSW Government in Australia has updated its procurement processes to address AI.[48] It considers four primary types of AI procurement: Acquiring a complete AI solution; "Hybrid development": Developing an AI system internally which involves procurement of specific components/ services from a vendor; Contracting a supplier to deliver a specific outcome, where the supplier leverages generative AI for the service provision; Change to a system/contract to incorporate AI functionality into pre-existing products or services.

Even without comprehensive regulations, vendor managers should expand internal risk frameworks to address AI-specific concerns and develop requirements for vendor contracts that protect the organisation.[49] It's crucial that procurement teams collaborate with data and AI specialists, legal experts, and business stakeholders to establish effective accountability guidelines, under each one of the above AI procurement types. A non-exhaustive list of concerns may include bias in training data, lack of transparency and explainability, restrictions on model and data usage rights, legal accountability for responses, copyright infringement, content ownership and intellectual property issues, data acquisition and transfer restrictions, and prevention of human exploitation. Procurement teams will be required to explicitly state as part of the contract which party is accountable for each risk, and in what stage does the accountability shift from vendor to the organisation and vice versa.

Contracts should require vendors to disclose how they detect and mitigate bias, their approaches to addressing potential harm in training data, copyright compliance measures, any implemented controls for output masking, as well as testing and benchmarking results in different contexts. Ideally, this information should be publicly available in the vendor's system cards.

[47] A law for foundation models: the EU AI Act can improve regulation for fairer competition
[48] IBM – AI Ethics and Governance in 2025
[49] Pondering AI: AI Literacy for All with Phaedra Boinodiris

Topic	Information required
General information	• Description of architecture • Input requirements (text, voice, etc) • Output format • Is it a stactic model?
Intended use information	• What are the use cases the model was developed for? • What use cases should be avoided?
Training information	• Description of training data (what is data source, multimodal, multilingual) • What data cleansing activities were conducted for training? • Context and task complexity • Description of the hardware and software the model was trained on and expectations for scaling into production
Performance information	• What evaluation metrics were used to validate performance? • Description of accuracy statistics
Ethics information	• Safety benchmarks used (BBQ, BOLD, Winogender, Winobias, RealToxicity and TruthfulQA) • Risks identified and mitigated (bias, generation of harmful content, intentional misuse, privacy)

Note: List not exhaustive

Source: Gartner (September 2024)

Managing vendor commitments extends beyond contractual compliance—models and systems require continuous monitoring for performance, drift, fairness, and quality. Even if vendors would meet their obligations to the fullest, once organisations leverage these models/systems for specific intent and in a specific context, they must ensure they comply with their own obligations to the regulators and internal policies.

The next section explores two vendor-provided system cards: OpenAI's "operator"[50] (an agentic AI model that may be leveraged as part of "hybrid development") and Microsoft Copilot 365[51] (a complete agentic AI solution). Given the above view of what is an "ideal" level of transparency and delineation of responsibility between providers and consumers, we struggle to form a satisfactory view of risks (OpenAI) and mitigations (Microsoft co-pilot). This may impact our ability to assess the use cases leveraging this model and system for risk, assess which tasks we may approve for development using this model/system, and lastly – have the assurance that across the AI value chain, each tier takes responsibility in a manner that supports the following tiers.

Example 1: OpenAI's Operator system card[33]

OpenAI's Operator is a Computer-Using Agent (CUA) model that can interact with graphical user interfaces autonomously.

OpenAI categorises the risks by user, model and resource:

Harmful tasks (a misaligned user): Users requesting the model to perform illicit activities, users attempting to use Operator for regulated activities without proper compliance, or users trying to engage in fraud, scams, or deception. OpenAI mitigations include clear usage policies that prohibit illicit activities, risk categorisation of tasks and actions, and models being trained to refuse harmful tasks.

Model mistakes (a misaligned model): Inadvertent actions causing harm (e.g., sending emails to wrong recipients), financial errors (e.g., purchasing incorrect items), or errors with potentially irreversible consequences. Mitigations include automated and human monitoring, including the requirement of user confirmation before finalising an action that affects "the state of the world", automatic refusal of high-risk tasks like banking transactions, and user supervision for sensitive operations.

Adversarial websites (a misaligned website/resource): Malicious instructions embedded in websites or content that could redirect the model, or third-party content tricking the model into performing unintended actions. Mitigations include website restrictions to prevent navigation to potentially harmful sites.

In addition, OpenAI discusses two frontier risks: The usage of Operator in biorisk tooling, and model autonomy (self-exfiltration, self-improvement, resource acquisition). General mitigations for frontier risks include prohibiting harmful and/or illicit tasks, pausing execution when user becomes inactive or navigates away. Operator was evaluated on biological design and model autonomy evaluation benchmarks. Operator scored low on both. OpenAI concluded that the model's autonomy was hindered by its visual input and cursor output modalities (i.e., making OCR mistakes and therefore cannot copy API keys or change code) to achieve full autonomy.

OpenAI states that they have implemented security defences against prompt injection and other security risks, assessing vulnerabilities using external and internal read teaming from multiple countries. They also mention that they have implemented mechanisms to rapidly update the model in response to new attacks.

[47] Operator System Card
[48] Copilot for Microsoft 365 Risk Assessment QuickStart Guide

Example 2: Microsoft 365 Copilot[34]

Microsoft 365 Copilot is an agentic AI SaaS, or in other words a black-box system. Users do not have access to its underlying training data, architecture nor system prompt. Rather, they can only view an output in response to their input. In general, Microsoft does not make any metrics available to alert issues with performance, quality, bias or drift of Microsoft 365 Copilot. As such, it is up to the organisation to extract input-output pairs and implement real-time monitoring of relevant metrics.

Bias - Microsoft notes that there are some measurements in place for fairness and voidance of stereotypes. In addition, as this is an AI system that leverages OpenAI's GPT-4 model, they refer their users to GPT-4's model card for more information on bias mitigation on the training data.

Disinformation/ungroundedness – Microsoft notes that responses are grounded in customer data to reduce the chance of hallucination using a combination of a RAG architecture and system prompting. It has groundedness, relevance, and similarity metrics in place to measure response quality. There is content filtering in place to reduce the chance of hallucinations. Microsoft notes that the responsibility lies with the end user ("user-in-the-loop"): "The application cannot spread disinformation on its own".

Harm – Microsoft employs abuse monitoring and content filtering. They note that there are multiple metrics in place to measure likelihood to produce hateful, violent, sexual and self-harm related content.

Overreliance on AI in decision making – Microsoft labels/informs users that they are interacting with AI, and it may be making mistakes. They do not mention any metrics in place to measure accuracy. As such, it is up to the organisation to implement policies and metrics for decision making.

Data leakage – There are data restrictions between users, teams, groups. Microsoft leverages the user identity-based access boundary so that the system will not present data that the user is not permitted to. Similarly, the grounding process only accesses content that the current user is authorised to access. Integration with 3rd party tools is controlled through the application. Microsoft notes that data governance is a shared responsibility. The effectiveness of their data access controls relies on the effectiveness of the implemented organisational data controls (access management, labelling of sensitive data, etc.), see section above for more information.

Intellectual property - Microsoft offers to its customers indemnity from IP infringement claims arising from the use and distribution of the output content generated by Copilot services.

Resilience and security – These elements of the risk report have the most detail. Microsoft notes that their implementation goes through rigorous testing and evaluation including red teaming for AI-specific security risks and 3rd party assessments. They include a list of pre-deployment security practices they follow to test their own application and OpenAI's code.

Explainability – Instead of noting how they help users understand how the system outputs were derived from the inputs, this section of the report covers what aspects of transparency should be met (intended use, high risk use, limitations) without the information itself.

Privacy – Organisation data is not shared, existing restrictions and access control apply to AI-infused systems. Organisational data remains private and is not used to train Microsoft's foundation models without permission. Data is not available to OpenAI. They have an unspecified metric in place to measure the likelihood of generating identified content.

Practical recommendations for AI procurement

1. Engage business use case owners, technical stakeholders, and gain understanding of which AI procurement type is planned.
2. Invest in the AI literacy of the procurement team and all stakeholders involved as decision makers or consumers of the procured AI.
3. Value addition and risk appetite: Identify how this 3rd party solution/component is to be leveraged across the business (across processes and in the ecosystem), how it fits into the business strategy (value orchestration), what are its planned use cases (approved/non-approved tasks).
4. Start the risk assessment early (see section above): Identify use case by use case risks and controls.
5. Include clauses into the contract which cover:
    A. Mitigation of any residual risks, explicitly stating which controls are to be owned by vendor.
    B. How performance of vendor's component/system may be assessed or monitored.
    C. How future features and/or system changes will be managed.
6. Assess vendor's ability to manage the risks identified.
7. Assign people in the organisation with the responsibility of monitoring the 3rd party component/system for performance and value.

Roles and Responsibilities in Managing Agentic AI

Successfully managing AI systems requires the collaboration of people in different roles across the organisation. As the risks amplify, the importance of a clear operation model and delineation of responsibilities becomes even more instrumental for the organisation's success. Existing organisational AI governance needs to be uplifted, and extended, to account for the amplified risk associated with agentic AI. Importantly, those who are accountable for AI outcomes need power and a funded mandate to perform their role. Often seen as a side responsibility, as organisations move to scaling AI, these roles become more of a full-time job.[45]

Sample Organisational responsibilities for Agentic AI[52]

Action	Accountable	Responsible	Consulted	Informed
Define organisational risk appetite for agentic AI	Executive level	Management level	Enterprise risk	Everyone
Define responsibilities, policies, guidelines to address current gaps in organisational readiness	Executive level	Management level	Enterprise risk	Everyone
Define the "should we" questions / traffic light system	Executive level	Enterprise risk	Management level	Everyone
Raise awareness and train the workforce in interacting with agentic systems	Executive level	Management level	Technical stakeholders	Everyone
Establish organisational structures and communication pipelines	Executive level	Management level	Enterprise risk	Everyone
Understand regulation & translate to business & technical requirements	Executive level	Enterprise risk	Management level	Everyone
Expand risk & control libraries to reflect new & amplified risks	Executive level	Enterprise risk	Management level, Technical stakeholders	Everyone
Keep track and implement protections against new adversarial attacks	Executive level	Technical stakeholders	Enterprise risk	Management level
Update the risk assessment process	Executive level	Enterprise risk	Management level, Technical stakeholders	Everyone
Sponsor new use cases and validate their suitability	Management level	Product/Service owners	Enterprise risk, Technical stakeholders	Everyone
Introduce new agentic AI systems to the registry and trigger their risk assessment	Management level	Product/Service owners	Enterprise risk, Technical stakeholders	Everyone
Introduce new models to the registry and assess which tasks and tools they are approved for	Technical stakeholders	Product/Service owners	Enterprise risk	Everyone
Oversee and manage the risk assessment process for models and systems	Management level	Enterprise risk	Technical stakeholders	Product/Service owners

[52] Understanding Responsibilities in AI

Action	Accountable	Responsible	Consulted	Informed
Suggest relevant controls for identified risks	Enterprise risk	Technical stakeholders	Product/Service owners	Everyone
Implement guardrail controls and guardian models	Management level	Technical stakeholders	Product/Service owners, Enterprise risk	Everyone
Implement real-time monitoring	Executive level	Technical stakeholders	Enterprise risk	Everyone
Design repeatable architectural patterns	Executive level	Technical stakeholders	Product/Service owners, Enterprise risk	Everyone
Identify risks of common architectural patterns	Enterprise risk	Technical stakeholders	Product/Service owners	Everyone
Define ethics by design, security by design guidelines for agentic AI	Executive level	Technical stakeholders	Enterprise risk, Management level	Everyone
Define and capture agents' intents and goals	Management level	Product/Service owners	Enterprise risk, Technical stakeholders	Everyone
Define constraints of action space (agents' do's and dont's)	Management level	Product/Service owners	Enterprise risk, Technical stakeholders	Everyone
Define which tasks/steps should be raised to a human for approval/oversight	Executive level	Management level	Enterprise risk, Technical stakeholders, Product/Service owners	Everyone
Human evaluation of results	Management level	Product/Service owners	Enterprise risk, Technical stakeholders	Executive level
Define least disruptive default behaviours of agents	Management level	Product/Service owners	Enterprise risk, Technical stakeholders	Executive level
Provide traceability - logging and audit capabilities - of agentic systems	Executive level	Technical stakeholders	Enterprise risk, Product/Service owners	Everyone
Provide explainability of agent actions	Management level	Technical stakeholders	Enterprise risk, Product/Service owners	Everyone
Implement reliable attribution of agent actions	Product/Service owners	Technical stakeholders	Enterprise risk	Everyone

Starting the Journey with Agentic AI

For organisations looking to incorporate agentic AI, we emphasise the need for a strategic, phased approach that balances innovation with risk management. This includes conducting thorough risk assessments, establishing clear governance structures, investing in talent development, and committing to ongoing monitoring and improvement. Additionally, it involves determining your strategy and approach to developing agents in the enterprise and effective planning to ensure the business value is achieved, the approach is scalable, and we can ensure trustworthy implementations.

When to Use AI Agents?

Broadly, the business value and associated complexity need to be assessed to determine whether to leverage agentic systems. Whilst you can achieve better performance, particularly when applied to complex use cases, the trade-off is often greater cost, latency[53] and potentially risk. Given this, the application of AI agents and agentic systems is most recommended where the complexity of the problem warrants it, where flexible and adaptive decision making is required and where the business value outweighs the cost.

The simplest solution to many business problems will likely be well designed orchestrations of traditional models and LLMs. These AI systems will provide greater predictability in outcomes for business processes that are well understood. As such, AI Agents will not be suitable for all business problems. The intended business outcomes to be delivered, and associated risk and return, needs to be well understood to design an effective AI system. However, AI Agents are perfect for complex business process execution; and specifically, a purposeful orchestration of AI Agents, LLMs, business services, tools and data in multi-agent systems will drive significant value in the enterprise.

Key insights to consider in the design of AI agents and agentic systems:[54]

- **User trust in outcomes is key:** rovide the ability to transparently understand and inspect the reasoning steps. How we design user interactions with AI agents is just as critical as the technical implementation to ensure value is realised.
- **AI Agents amplify, and introduce new, risks to design for:** ensure effective risk identification and mitigation design up front in the system design. Guiding AI behaviour does not guarantee strict adherence to rules. Comprehensive enforcement mechanisms are required for AI agents. Evaluation and real-time monitoring are critical.
- **Focus on one persona at a time:** Determine the most valuable persona to ensure impact. Iterate and refine based on user needs, usage, performance data, monitoring outcomes and business outcomes.
- **There is no one size fits all:** No single AI agent design, or framework, is definitive. Right fit the design to your problem. There are rapidly evolving agentic frameworks, models and orchestration tools to accelerate the build of agentic systems. These however must be assessed to ensure tailoring to your specific business problem and acceptance criteria.

Common pitfalls in designing AI agents and how to resolve:

- **Planning:** Poorly defined tasks and personas create inadequate planning mechanisms with lead to agents struggling with complex tasks
  - Resolution: Design detailed personas, goals, constraints, and expected outcomes for each agent. Use task decomposition, multi-plan selection, and reflection to improve planning
- **Efficiency:** AI Agents can get stuck in infinite loops without goal progression or can have tool calling failures where tools are misused, or outputs are misinterpreted
  - Resolution: Define clear tool parameters and validate outputs with verification layers. Define clear termination conditions and consider enhancing fault tolerance where appropriate to improve error handling
- **Scalability:** Increased workloads and complexity can cause resource intensive costs to run or inefficiency
  - Resolution: Optimise context length, minimise API calls, and leverage smaller models where appropriate. Integrate dynamic resource management capabilities.
- **Reasoning:** AI Agents can be prone to unpredictable, or inconsistent, behaviour and can struggle with nuanced judgement.
  - Resolution: Enhance reasoning through prompting techniques, fine-tuned or more capable models. Where required enforce strict inputs and output validations to guide agents
- **Safety:** Limited auditability and transparency in agent actions, which hinder accountability and triaging.
  - Resolution: Implement effective guardrails, action constraints, escalation protocols and feedback loops. Ensure effective, and timely, monitoring and evaluation for adherence to standards

[53] Anthropic: Building effective agents
[54] Hard-Earned Lessons from a Year of Building AI Agents | by Maya Murad | Feb, 2025 | Medium

Practical Steps to Get Started with AI Agents

1. Determine your strategy & plan

a. Define your strategy in determining the use of AI agents and how this will work with your current people, tools and processes
b. Identify business value for key potential use cases and applications
c. Define detailed personas, goals, constraints, and expected outcomes for identified use cases and workflows at the most granular level
d. Implement the use case and its controls at the same time
e. Evaluate use case and controls performance prior to deployment, including FinOps KPIs

2. Ensure safety & trust

a. Define your AI ethics principles and determine the kind of relationship your company wants to have with AI
b. Define clear functional and non-functional requirements to operationalise your AI principles
c. Uplift your risk and controls library for Agentic AI
d. Establish a centralised, standardised repository of use cases, consumable to both technical and nontechnical stakeholders
e. Stand up a cross-organisational AI literacy program
f. Ensure comprehensive end-to-end risk management processes. Shift the risk assessment process left, strive to automate as much of it as possible
g. Incentivise a risk-oriented innovation culture, with easy-to-use risk identification guidance, and an operating model with clear accountabilities
h. Implement real-time monitoring
i. For each use case, identify risks to govern the AI agents and/or agentic systems, and design effective controls to mitigate. Determine the level of rigour required to mitigate for each component of the agentic system

3. Determine approach to scalability

a. Determine the strategy and approach to building AI agents leveraging agent frameworks to accelerate scale.[55] Whilst agentic frameworks help accelerate development, they may also potentially introduce unnecessary complexity and abstractions
b. Determine the approach to enabling access to the frameworks and tools for business domains, developers and product teams
c. Start small, prove, and refine

Conclusion

The emergence of agentic AI, and AI agents, in financial services presents both exciting opportunities and unique challenges. By understanding the capabilities and limitations of these advanced AI agents, financial institutions can harness their potential while critically understanding and mitigating associated risks. AI agents represent a significant opportunity to enable dynamic response to a continuously changing business environment. Through purposeful strategic planning, robust risk management, and a commitment to responsible AI practices, the financial sector can successfully navigate this new frontier to transform customer experiences and realise business value. It will be increasingly critical for risk functions to consider how we, at scale, apply and manage effective controls to our people, processes, and technology. In doing so, we can move in tandem with business innovation, and ensure the potential value is realised from this opportunity, whilst critically managing the risks.

[55] AI agent frameworks: Choosing the right foundation for your business

About the authors

Michal Chorev, AI Governance Lead, IBM Consulting Australia

Michal is the AI Governance Lead for IBM Consulting Australia, where she advises clients across all stages of AI adoption, from strategy through implementation and responsible scaling. She guides organisations on establishing ethical AI frameworks, governance structures, and risk management approaches to ensure responsible and compliant AI deployment. Previously at IBM Research, Michal led the successful adoption of AI-based decision support systems for Healthcare and Pharma clients, focusing on explainability, causal inference, and multi-modal prediction capabilities. With over 15 years of experience leading R&D teams in AI and advanced analytics, she brings deep technical expertise combined with practical implementation knowledge. Today, Michal advises clients across multiple industry sectors on how to design and implement trustworthy AI systems that deliver exceptional business value while maintaining ethical standards and regulatory compliance.

Richie Paul, Generative AI and Strategy and Transformation Lead, IBM Consulting Australia

Richie leads IBM's Generative AI Practice where he advises IBM clients on the adoption of Generative AI, the business and technology architectures for transformation, new enterprise strategies, and modernisation initiatives. He is also an advisor to multiple Australian government departments and a contributor to government digital taskforces, including the NSW Government's AI Taskforce. As a strategy consultant, Richie has over 20 years of experience implementing transformational digital business solutions working in multiple industry sectors across Asia Pacific and Europe.

Joe Royle, AI Strategy Lead, IBM Consulting Australia

Joe is the AI Strategy Lead for IBM Australia, where he advises clients in the Banking and Financial Services sector on the responsible design and deployment of AI products, platforms and services. He specialises in enterprise strategy, AI-led business and service design, with deep expertise in designing and scaling AI-enabled digital platforms. Joe has spent his career leading complex digital transformation programs, with a focus on delivering AI solutions that are governable, ethical and aligned to organisational strategy. He is a strategic advisor to clients on AI risk, governance and product/platform design, and plays a leadership role in IBM's design practice, helping organisations adopt AI at scale in a responsible and sustainable way.

With special thanks

Kasia Ligertwood, Senior Manager, Artificial Intelligence Tech Risk, Commonwealth Bank of Australia

Kasia Ligertwood brings specialised expertise in applying risk frameworks to emerging technologies in banking, built through her experience across risk assurance consulting, internal audit and second line technology risk. Now embedded in CBA's AI Technology Risk team, Kasia provides valuable insights into the distinct risk profile posed by agentic AI. She champions responsible AI deployment through pragmatic, tailored control frameworks that balance innovation with the unique regulatory and operational realities of financial institutions.

Alejandro Eizagaechevarria, Executive Manager, Artificial Intelligence, Technology Risk and Analytics, Commonwealth Bank of Australia

Alejandro is the Executive Manager of Artificial Intelligence Technology Risk at Commonwealth Bank of Australia, where he partners with business and technology teams to ensure the safe and responsible adoption of AI across the Bank. With over 15 years of experience in risk management and technology audit within the financial services sector, Alejandro brings extensive expertise in Technology, Data, and Artificial Intelligence risk. In his current role, Alejandro leads a team of risk Subject Matter Experts (SMEs) focused on AI, driving the Bank's initiatives in AI governance and AI risk management to ensure AI adoption aligns with the Bank's risk appetite, regulatory obligations, and ethical standards.

Sam Gandy, General Manager, AI and Data Risk, Commonwealth Bank of Australia

Sam Gandy is the General Manager of AI & Data Risk at Commonwealth Bank, where he leads enterprise-wide strategies to identify, assess, and mitigate risks associated with data and artificial intelligence. With a background in technology architecture, Sam brings a unique perspective to the intersection of technology, and risk management. In his current role, Sam is instrumental in shaping the bank's approach to AI governance, ensuring that innovation aligns with risk appetite, regulatory requirements and ethical standards. His work focuses on assuring robust frameworks that balance technological advancement with risk mitigation, fostering trust and resilience in the digital landscape.

Phaedra Boinodiris, Global Leader for Trustworthy AI, IBM Consulting

Phaedra, a Fellow of the London-based Royal Society of Arts, has been dedicated to promoting inclusion in technology since 1999. She currently leads IBM Consulting's Responsible AI Practice, which she co-founded, and serves as a member of the Global Council for Responsible AI. She is also the author of AI for the Rest of Us. In addition, Boinodiris co-founded the Future World Alliance, a 501(c)(3) organiszation committed to curating K-12 education in AI ethics. Her accolades include being named the 2030 Responsible AI Leader of the Year, receiving the United Nations Woman of Influence in STEM and Inclusivity Award, and being recogniszed as one of the Top 100 Women in the Games Industry.

Matthew Bellio, Senior Data Science Manager, IBM Consulting

Matt is one of the Global Technical Leaders for the AI Strategy & Governance offering at IBM Consulting, where he drives thought leadership and client enablement around the trustworthy design, deployment, and oversight of Predictive, Generative, and Agentic AI systems. With over a decade of hands-on experience as a data scientist, engineer, and architect, across diverse industries, he prides himself on helping clients navigate all aspects of the AI lifecycle responsibly— bridging innovation with practical, sustainable outcomes tied to business value.