Centre for Information Policy Leadership
HUNTON

# Agentic AI

## Fostering Responsible and Beneficial Development and Adoption

November 2025

## About the Centre for Information Policy Leadership

The Centre for Information Policy Leadership (CIPL) is a global privacy and data policy think tank within the Hunton law firm that is financially supported by the firm, 85+ member companies that are leaders in key sectors of the global economy, and other private and public sector stakeholders through consulting and advisory projects. CIPL's mission is to engage in thought leadership and develop best practices for the responsible and beneficial use of data in the modern information age. CIPL's work facilitates constructive engagement between business leaders, data governance and security professionals, regulators, and policymakers around the world. Nothing in this document should be construed as representing the views of any individual CIPL member company or Hunton. This document is not designed to be and should not be taken as legal advice. For more information, please see CIPL's website at:

**http://www.informationpolicycentre.com/**

CIPL would like to thank HCLTech, Salesforce, and Workday for their support for and thought leadership contributions to this report, as well as all its members who shared additional insights and case studies.

HCLTech

salesforce

workday.

CIPL
Centre for Information Policy Leadership
HUNTON

**DC Office**
2200 Pennsylvania Avenue
Washington, DC 20037
+1 202 955 1563

**London Office**
30 St Mary Axe
London EC3A 8EP
+44 20 7220 5700

**Brussels Office**
Avenue des Arts 47-49
1000 Brussels
+32 2 643 58 00

# Table of Contents

# Introduction

Agentic artificial intelligence (agentic AI) has the potential to transform business processes and customer experience across a range of industries. By one estimate, over 40 percent of enterprise workloads will utilize autonomous agents by 2027,[1] and by 2028, as much as 15 percent of day-to-day work decisions will be made autonomously through agentic AI.[2] To reach these goals, companies are working to rapidly scale their investment, development, and adoption of agentic AI to help better serve customers and reap benefits, such as greater productivity and efficiency, optimization of organizational resources, and personalization of products and services.

In recent years, business leaders, policymakers, and regulators have all recognized and sought to address a range of legal, governance, and operational challenges associated with AI, many of which will be amplified by agentic AI. Leading organizations adopting agentic AI are striving to develop governance solutions to challenges associated with system autonomy, privacy, security, and transparency, among other new challenges.

The Centre for Information Policy Leadership (CIPL) has long promoted organizational accountability and a risk-based approach as the cornerstones of trustworthy, responsible development and deployment of AI technologies.[3] This new report on *Agentic AI: Fostering Responsible and Beneficial Development and Adoption* builds on that work by highlighting both the opportunities and challenges posed by agentic AI. The report outlines key benefits, risks, mitigation measures, relevant tensions with data protection, and recommendations for both organizations and regulators. The paper also features current best practices and case studies from companies that have demonstrated a commitment to responsible development and adoption of agentic AI.

This current report is part of CIPL's broader Agentic AI Project, which aims to highlight agentic AI's benefits, clarify the potential risks, and identify practical solutions for business leaders and policymakers to address them. This first report chiefly focuses on business-to-business (B2B) applications of agentic AI; subsequent research will delve more deeply into business-to-consumer (B2C) applications.

Implementing and demonstrating accountability in technology innovation are continuous processes, and we anticipate that the findings and organizational practices presented in this report will evolve as the technology matures and organizations adapt alongside it. We hope that this report will help demystify agentic AI technology, identify potential governance challenges and solutions, and foster discussion and cooperation that promotes the continued development and deployment of responsible and beneficial agentic AI.

CIPL would like to thank HCLTech, Salesforce, and Workday for their support for and thought leadership contributions to this report, as well as all its members who shared additional insights and case studies.

# Defining Agentic AI

An agentic AI system is characterized by four core capabilities:

1. **Autonomy** It can move beyond rule-based automation to perform tasks on its own without constant human oversight or intervention, based on need. For example, an autonomous wealth management system can monitor market developments and undertake transactions to rebalance a portfolio to meet specified parameters, without the need for constant human intervention.
2. **Adaptability** It can adjust its behavior in real-time based on changing conditions and environments. For example, an agentic online learning application can adjust lesson materials in response to continuous signals as to how quickly students are mastering or struggling with specific concepts.
3. **Learning** It can adapt from its interactions and experiences to improve future decisions and performance. For example, banks can employ agentic fraud detection systems that constantly learn from transaction data to recognize changes in fraud patterns.
4. **Orchestration** It can coordinate individual agents and manage multiple tasks to achieve complex, broader goals and execute end-to-end workflows. For example, for supply chain management, an agentic system can coordinate across a range of operational individual procurement, inventory management, transport, and logistics providers to enable seamless operations.

Some agentic AI systems are built on top of foundation models, such as generative AI (genAI) models, leveraging their core intelligence capabilities (i.e., reasoning, multitasking ability, understanding and generation of content) to interpret, break down, and execute high-level tasks. Agentic AI systems can also work across third-party applications and utilize a complex network of individual agents, each designed with specific goals and abilities, but capable of working collaboratively with other agents within the broader agentic system to tackle complex tasks. Over time, agentic AI systems adapt from their interactions and refine their outputs to make processes even more efficient.

It is also important to define the distinction between an **agentic AI system** and **AI agents** because it informs the way both organizations and policymakers identify and mitigate potential risks.

Many of the most complex and systemic risks are not necessarily relevant to individual agents acting in isolation, but to multi-agent systems (or an agentic AI system), where multiple agents are interacting with each other and various environments.

We can compare the relationship and distinction between agentic AI systems and AI agents to those between an orchestra conductor and the individual musicians. The conductor acts in line with the broader objective of

delivering a great musical experience, directing individual musicians and setting the overall tone, timing, and flow of the performance. Similarly, the agentic AI system works towards an overarching desired outcome, coordinating multiple AI agents, adapting to real-time data, and handling complex tasks by exercising its autonomy and reactive decision-making. On the other hand, each AI agent is optimized and built to automate specific tasks within defined parameters, just as individual musicians are highly skilled at playing their own instruments. To provide an example, an organization might employ an agentic AI system for supply chain management that orchestrates individual AI agents to complete specific tasks (e.g., procurement, transportation planning) in a coordinated manner. Agents will act according to the instructions of the agentic system or the user, and while capable on their own, they can deliver greater impacts and benefits when orchestrated as parts of a larger agentic system.

> **CASE STUDY 1**
>
> Workday published an "AI Masterclass" to help companies learn to responsibly build, deploy, and govern AI for business. This Masterclass includes a chapter dedicated to understanding AI Agents and their potential in the enterprise context, walks through different types of agents, and shares strategies for successful AI deployment.[4]

# Value Adds from B2B Agentic AI Development and Deployment

The development and deployment of agentic AI systems at the enterprise level offers a wide range of strategic, operational, and economic benefits.

Organizations cite the following as the greatest value-adds from their investment into agentic AI in the B2B context:

## 1. Efficiency and Productivity

Agentic AI can autonomously handle routine tasks, monitor processes, and initiate actions without continuous human prompting. By deploying agents to automate repetitive workflows that may have previously required complex, hard-coded decision trees and optimizing resource allocation between human and AI agents, agentic AI can reduce labor costs, minimize inefficiencies, and enable employees to focus on higher-value, strategic work. This can overall help organizations streamline operations, reduce operational bottlenecks, cut down on manual errors, and achieve faster turnaround times, which can, in turn, directly improve the bottom line.

### CASE STUDY 2

Financial institutions are exploring the use of AI agents to combat financial crime and streamline "anti-money laundering" (AML) and "Know Your Customer" (KYC) processes. Individual agents can be assigned different tasks (e.g., for monitoring or troubleshooting the data pipeline, researching and analyzing data from various sources, reviewing workflow outputs and ensuring quality), and organizations are predicting significantly increased quality and consistency of AML/KYC outputs, while reducing the need for manual intervention and overall operational expenditures, particularly compared to traditional KYC/AML methods.[5]

### CASE STUDY 3

Stanford Health Care, in partnership with Microsoft, is building a "healthcare agent orchestrator" that can support oncologists prepare for tumor board presentations, a meeting where a team of specialists (e.g., oncologist, radiologist, pathologist, surgeon) comes together to discuss a care plan for individual patients.[6] Preparing for these presentations has always been a necessary, yet immense burden on providers' time and capacity. However, AI agents can help reduce this burden by taking on some of the more time-consuming components of documentation (e.g., building a patient timeline, synthesizing current medical literature, identifying relevant clinical trials), leaving more time for providers to spend with their patients. The orchestrator can analyze and reason over diverse types of data, including imaging, genomic data, and clinical notes, and integrate with existing workflows housed in non-agentic applications (e.g., word processing and presentation software). Although the application is still in the testing phases and has not yet been used in clinical settings, early testers predict these agents will be extremely impactful in their day-to-day schedules and are eager to deploy.

**CASE STUDY 4**

HCLTech's AI Force is a patented service transformation platform designed to enhance efficiency across the software engineering and IT operations lifecycle.[7] It integrates AI and genAI technologies, including commercial options (e.g., Azure, OpenAI, and Google Gemini), as well as open-source models. The platform automates routine tasks, accelerating processes, such as testing (by 35 to 50 percent) and modernization (by 50 to 60 percent). It also includes input and output security scanners to help ensure safe and responsible AI usage.

## 2. Adaptability and Scalability

Agentic AI systems process vast data streams in real time, providing actionable insights for smarter decisions. By detecting patterns and forecasting outcomes, they enable businesses to make data-driven choices with confidence and to respond more rapidly to market changes, regulatory updates, or operational disruptions. Agentic AI also enables businesses to create more personalized and engaging interactions for users by mimicking human-like decision-making and user interactions and adapting to customer preferences and behaviors.[8] These capabilities can help increase customer engagement, satisfaction, and loyalty without increasing manual workload for employees.

**CASE STUDY 5**

Mastercard recently unveiled its new agentic payment program, "Agent Pay," delivering agentic payments technologies that will integrate seamless payment experiences with tailored, data-driven recommendations and insights provided on conversational platforms for individual users and businesses.[9] The program integrates "Agentic Tokens," building upon existing tokenization capabilities to enable initiation of payments in a trusted, secure way and strong payment authentication solutions, including Mastercard Payment Passkeys, verifiable tracking of consumer purchasing intent, as well as fraud and cybersecurity solutions to support safe and secure agentic commerce.

**CASE STUDY 6**

Mercedes-Benz will be one of the first automakers to implement Google Cloud's "Automotive AI Agent," specifically integrated with Google Maps and designed for multilingual support to deliver highly personalized in-car agents.[10] Drivers will be able to engage in natural language conversations and receive personalized recommendations about topics, such as nearby points of interest, traffic conditions, and directions. In addition to driver-oriented functions, AI agents can be used to enhance fleet management through predictive maintenance and tools to optimize fuel use.[11]

Agentic AI can also support business functions in scaling their internal processes. For example, agentic AI can streamline privacy compliance by gathering information on data flows and data mapping, executing transparency and privacy notices to individuals, enabling data security, assisting in exercise and response to individuals' rights requests, and monitoring and reporting activities.

**CASE STUDY 7**

Salesforce's Agentforce, the agentic layer of the Salesforce Platform, allows businesses to configure agents with strict boundaries around the processing of personal data, prioritize data minimization, and enable compliance with global privacy regulations.[13] Salesforce actively encourages businesses to implement sufficient controls to guide agent behavior through the entire customer lifecycle. This can include configuring agents to present the business's privacy statement during the first point of contact, better understand how these policies impact them, and provide "just-in-time" notices if additional personal data is collected.

## 3. Decision Quality and Consistency

With proper guardrails and when developed or deployed responsibly, agentic AI systems can minimize the risk of human error, variability, and even misconduct by applying standardized logic across a variety of tasks. This can be particularly valuable in contexts that require consistent application of organizational or regulatory policies. For example, agentic AI can be deployed to automate due diligence requirements, such as know-your-customer (KYC) and anti-money laundering (AML) or monitor for instances of potential internal policy violations or regulatory noncompliance.

### CASE STUDY 8

SAP has launched a series of AI agents called "Joule Agents" that can orchestrate networks of agents across an organization to complete tasks in areas, such as supply chain management, marketing, and finance. For example, Joule can be used to identify errors more rapidly in invoices to reduce the frequency of disputes and resolve them more quickly.[12]

### CASE STUDY 9

An AI agent developed by Castellum.ai automates KYC by screening records for potential concerns and taking follow-up actions. For example, if the tool identifies a proof-of-address that appears altered, it automatically asks the provider of the record for clarification.[13]

## 4. Innovation and Problem-Solving

Agentic AI can foster innovation by augmenting employee capabilities rather than replacing them and enabling teams to focus on high-impact, strategic problem-solving and creative thinking. Organizations can leverage agentic capabilities, such as automating routine tasks, providing recommendations based on learning, and generating ideas, to encourage experimentation and iteration across various business functions, including product development, operations, and strategy. Agentic AI systems can also expand the potential of vertical use cases of genAI tools, enabling LLMs with additional components and capabilities to move from reactive, passive tools to proactive, autonomous support.

### CASE STUDY 10

Workday's agents help enterprises with tasks like contract review (Contract Intelligence Agent) and personalized matching of internal talent to open roles (Talent Mobility Agent). Workday's Contract Intelligence, powered by Evisort, provides industry-leading contract AI to surface risks and opportunities, streamline access to insights, and enhance collaboration between legal and business teams.[14] It delivers a unified, intelligent contract repository with a wealth of structured data points from the contracts and related documents that are the lifeblood of organizations. This provides complete visibility into agreements without costly and detailed manual review. Not only can legal teams respond to contract-related inquiries promptly and efficiently, but they can also provide self-service access to actionable insights and share valuable contract data with teams across the enterprise—including procurement, finance, HR, compliance, M&A, InfoSec, and sales and revenue operations.

## 5. Resilience and Continuity

Agentic AI can act as an intelligent, autonomous layer to enhance business resilience. By providing continuous operational coverage to detect anomalies in the supply chain, automating crisis response systems, and predicting and proactively managing emerging risks (e.g., supplier health, weather patterns, market and geopolitical trends), agentic AI can safeguard critical operational functions and strengthen organizational preparedness.

> **CASE STUDY 11**
>
> Salesforce's "Agentforce for Consumer Goods" is a specialized solution that uses a domain-specific agent, including pre-build and custom AI agent actions for consumer goods companies, to monitor and enhance key operations, such as live inventory management, asset service, and customer support. The agentic AI system integrates with existing Salesforce and external systems to process data in real time and provide actionable recommendations, such as assessing the condition of existing inventory and recommending adjustments. By spotting issues early, "Agentforce for Consumer Goods" can help reduce the likelihood of inventory shortages, associated costs, and missed retail opportunities.

# Governance and Compliance Challenges and Solutions

While agentic AI's potential benefits are diverse and significant, the technology also carries challenges for governance and compliance that will be important to address to cultivate trust among businesses, consumers, and regulators and support broader adoption. In this section, we highlight examples of these challenges and paths to address them, while recognizing that additional challenges, solutions, and benefits are likely to become apparent as adoption accelerates.

Most of these challenges are not new to agentic AI: many emerged with the introduction of AI technology and were later amplified by genAI. However, agentic AI's broadened capacity for autonomous decision-making without continuous human oversight can heighten these existing challenges in scale, impact, and complexity. Fortunately, agentic AI also offers new potential for addressing these challenges. We discuss them below, as follows:

- Purpose and Proportionality
- Control and Accountability
- Integrity and Reliability
- Transparency and Explainability
- Security and Resilience
- Alignment and Human Agency

## Purpose and Proportionality

Agentic AI training and deployment raise considerations with respect to core data protection principles, including the legal basis for processing, purpose definition and limitation, proportionality, data minimization, and more. We address these issues here (along with several related to non-personal data).

### Legal Basis for Processing

In many jurisdictions, the collection of personal data may require a legal basis, such as consent, contractual necessity, public interest, or the legitimate interests of the controller or a third party. In jurisdictions that do not specify legal bases, data collection must frequently still follow other requirements, such as purpose specification and fairness, to be lawful.

Obtaining meaningful consent in agentic AI systems, where downstream processing of data may not be foreseeable by users or developers, can be quite challenging and overly burdensome. Enabling responsible data processing practices through robust data governance (see more in *"Adopting Robust Data Governance Measures and*

*Guardrails"*) will be crucial to demonstrate accuracy and reduce the risk of goal misalignment, or situations where agents pursue objectives in ways that conflict with user intent, organizational values, or even legal obligations. This risk can be amplified in agentic AI systems, which can rapidly scale misaligned actions, such as spamming or discriminatory outcomes.

Because AI agents are capable of rapidly processing such decisions at greater scale, human oversight will be a key guardrail in mitigating this risk. High-risk, consequential decisions should be accompanied by robust data governance processes and human involvement, aligning with GDPR Art. 22 (see more in *"Human disempowerment and loss of autonomy"*).[15] Developers and deployers of agentic AI systems should also carefully assess decision making processes with respect to recent legal cases, such as *OQ v Land Hessen* (i.e., "Schufa decision").[16]

## Purpose Definition and Limitation

Agentic AI systems require vast amounts of data and may self-initiate data collection from multiple sources to accomplish their goals, including datasets that were not originally foreseen at the time of system design or deployment. Agents may also need to collect new data or repurpose existing data to optimize performance, which may go beyond the original scope of the intended purpose. This can raise questions about whether the data collected was necessary and collected for a specific, lawful purpose, particularly as agents may need to evolve their goals in real time.

Because agentic AI systems operate in dynamic and often unpredictable environments, they must be able to adapt their data use to changing circumstances and emerging objectives. However, rigid interpretations of the purpose limitation principle could inhibit lawful and beneficial innovation in agentic AI. Thus, legal frameworks should recognize this need for flexibility by allowing organizations to define processing purposes broadly enough to accommodate reasonable evolution in the system's functions and objectives, while ensuring that data is used in ways that reflect people's reasonable expectations.

## Proportionality and Data Minimization

Some agentic AI applications may require access to data from a wide breadth of sources, including personal data, to complete complex tasks. Potential privacy risks in such use cases include:

1. **Overcollection of data in pursuit of achieving a goal** An agentic AI system is designed to operate autonomously to achieve specific goals, often by dynamically determining what data is needed and how to obtain it. However, in doing so, it may inadvertently over-collect data, gathering more than what is necessary to complete its intended objective, disproportionate to the legitimate business need. In the B2B context, this could include pulling full customer records when only a small subset of data is required or logging large volumes of data to memory even after the task is complete. Overcollection can also increase points of vulnerability for unauthorized access to information, or lead to inadvertent processing of personal data or retention of data beyond intended periods.[17]
2. **Over-profiling** Because agentic AI is designed to continue learning to improve performance, it could potentially create overly detailed profiles of individuals or entities that include sensitive information, such as medical information, financial status, and more. Without proper safeguards, such profiles could lead to biased decision-making or even serve as potential targets for bad actors.

CIPL has argued that data minimization should be understood and applied in a contextual and flexible manner, particularly in the context of AI. It also should not prevent the collection and use of data that is necessary and appropriate for achieving a high-quality outcome, whether that is a robust AI model or optimization of complex tasks or processes.[18]

It is important to consider that the lawfulness of data processing for AI under relevant laws, such as the GDPR, is often determined not by the sheer volume of data, but by the necessity of that data for a specified legitimate interest and a careful balancing of that interest against individual rights and freedoms. A landmark decision in summary proceedings by the German Higher Regional Court of Cologne, for instance, affirmed that large-scale processing of user data specifically for AI training can be justified under legitimate interests.[19] The Court accepted that alternatives, such as fully anonymized or synthetic data, may be insufficient for developing high-quality, regionally-nuanced AI, and crucially, that the "necessity" of the data must be assessed for the dataset as a whole, rather than on an impractical per-datum basis.

Taking a nuanced approach to the data minimization principle can help address the tension between data protection and the realities of modern AI development. Relevant legal frameworks, such as the EU AI Act,[20] have already recognized this approach as a potentially valid pathway to reconcile innovation with robust data protection.[21]

A potential technical solution to further reconcile the data minimization principle may include limiting agents' access to only the data necessary for specific tasks or segmenting data stores (e.g., by purpose or sensitivity), so that access is granted only when a task's purpose aligns with the data's permitted use.[22] For example, an agent tasked with calendar scheduling would not be permitted to access health records, even if the datasets are available in the same environment, like a phone.

### Sensitive Information

CIPL has described why the processing of sensitive data can be valuable to ensure that AI applications function effectively, fairly, and safe.[23] At the same time, the risk of over-profiling (see more in *"Proportionality and Data Minimization"*) points to the importance of ensuring that agents use sensitive personal data appropriately and with well-considered limits. Similarly, the cybersecurity risks associated with agentic AI (see more in *"Security and Resilience"*) may be accentuated in circumstances where sensitive data may be needed to deliver the promised enterprise benefits, but the available security controls may not be sufficient to meaningfully mitigate the risks. This is why taking a risk-based approach in deployment, with guardrails and mitigations tailored to the level of risk, is critical, particularly in deployment scenarios involving the use of sensitive data (see more in *"Accountability and a Risk-Based Approach as Enablers of Responsible Innovation"* and *"Implementing a Comprehensive, Integrated Risk Assessment Process"*). It is especially important that controls be set so as to avoid "over-permissioning" access to sensitive data and ensuring that its use is targeted for appropriate contexts.[24]

It is important to note that sensitive data include "protected characteristics" under the GDPR and similar laws, such as race and ethnic origin, religious beliefs, political opinions, financial data, and medical information. It is equally important for agents to be deployed with proper guardrails to guide interactions with sensitive non-personal data, such as trade secrets or data otherwise subject to intellectual property protections. These data may also have legitimate and important uses in agent interactions within and between organizations, but such uses must be governed with care.

### Derived Use and Reuse

Agentic AI often processes large volumes of data from multiple sources to detect patterns, trends, and relationships and makes inferences based on that data. While this can generate valuable insights for businesses, it can also raise privacy risks by increasing the risk of inferring sensitive information (e.g., financial status or health conditions) about individuals or entities from seemingly non-sensitive data points (e.g., purchasing history, online activity). Similarly, agentic systems raise re-identification risks through the potential tracing of previously anonymized or pseudonymized data back to individual users. These risks are particularly relevant for agentic AI because these systems may autonomously combine multiple datasets across multiple teams, businesses, or third-party sources; retain and accumulate data over time, especially in the absence of explicit retention policies; and make decisions based on the inferred data.

### Data Sharing and Cross-Border Transfers and Onward Sharing

Agents raise novel questions around governance of the sharing of data across organizations. Organizations have long maintained agreements to govern the sharing of data with other organizations, as well as technical and operational controls to manage such exchanges securely. They will need to adapt these mechanisms for agent-to-human and agent-to-agent interactions. Data transfers across national borders raise additional considerations: many countries have already established rules around transfers of personal data to specific other countries, but data transfers enabled by agents will need to be accompanied by controls designed to ensure compliance with such rules.

### How Agentic AI Can Serve as a Privacy Enabler

While agentic AI poses potential privacy risks, many organizations recognize that AI agents can also serve as privacy enablers:

- Agents can be leveraged to fortify organizations' privacy efforts by acting as intermediaries and as the first line of defense. They can limit unnecessary human access to data, prevent excessive data collection and access from humans, and reduce the overall risk of privacy incidents.[25]
- Organizations are already using AI agents to automate and scale their privacy programs and controls, in ways that would be impossible without larger human teams. For example, organizations are using agents to:
  - Respond to data subject access requests (DSARs) in a timely manner;
  - Help triage incidents, quickly gathering relevant information to meet the request, and automating other historically time-consuming tasks; and
  - Reinforce existing privacy programs, while enabling human employees to have more time for other tasks.

While deployment of AI agents—or human agents—is never risk-free, organizations can proactively reduce risks by putting in place appropriate human controls to monitor for early signs of potentially risky behaviors. Technical controls can also play an important role in maximizing the benefits of agentic AI, while mitigating potential privacy risks across the AI lifecycle.[26] For example, techniques, such as Retrieval-Augmented Generation (RAG), can help minimize the amount of personal data needed for tasks, automating the process and providing organizations with higher quality, more accurate outputs. In addition, agents can be configured to provide privacy-related information in real-time to people using the agents, including interactive guidance on organizations' privacy policies and information on how to exercise data subject rights[27] Finally, organizations are implementing a range of privacy-enhancing technologies (PETs) to enhance privacy during AI model training and deployment, including synthetic data, differential privacy, and homomorphic encryption.[28]

## Control and Accountability

As agentic AI systems gain increased autonomy, expand capacity for decision-making, and interact with one another, previously understood definitions of roles and responsibilities will become increasingly complex. Key questions, such as who will be responsible for what across the AI value chain and how to implement meaningful human oversight and redress mechanisms, will need to be explored further between relevant stakeholders.

### Role Allocation Across the Lifecycle

The challenge of assigning responsibility, already made complex by the introduction of general-purpose AI, will only be amplified in the context of agentic AI. These systems are designed to make independent decisions that go beyond direct human prompting. This raises questions about who is ultimately accountable for their actions, not just regarding the processing of data, but also in the event of an unintended outcome that causes material harm. Should accountability lie with the entity closest to the generation of the harm? What is the role of human oversight in exercising control over agents' actions?

In the absence of specific regulatory guidance on agentic AI, organizations should adopt accountability and governance measures to enable adequate human oversight. Both developers and deployers have parallel, yet distinct roles to play in mitigating the risk of harms along the agentic AI lifecycle, as the various entities in the agentic AI value chain have different levels of control over, and visibility into, agents at various points of the lifecycle. For example, developers generally have control over how agentic AI systems are designed and trained, and how they interact with other software provided by the developer. Deployers, on the other hand, can control how agents are configured and used in practice, and have visibility into agent operations and underlying data that developers do not. These differences mean effective agentic AI governance requires effective collaboration and partnership between developers and deployers around a shared responsibility for AI risk management.

Below, we have outlined some examples of potential accountability controls for both developers and deployers, while recognizing that the appropriate design and allocation of controls are likely to vary according to the specifics of deployment scenarios (e.g., when multiple agents are interacting along the value chain). Regulatory guidance and contractual arrangements between service providers and deployers can also aid in clarifying the allocation of responsibilities for agentic AI systems across the value chain between developers and deployers.

| Developers | Deployers |
|---|---|
| Work at the system level (or, as the OECD calls it, "in the lab"[29]) to enable adequate risk mitigation through techniques, such as prompt engineering and testing and validating the system to return desired, appropriate outputs.[30]<br><br>Conduct AI red teaming to stress test agentic AI systems by simulating threat attacks from malicious actors, provoking unintended behaviors, and identifying how agents might exploit tools or deviate from intended goals.<br><br>Provide customers with information they need to adopt agentic AI products and services in a responsible manner and that outline both the capabilities and limitations of the technology and actions that deployers can take to further mitigate risks as the operators of the system—for example:<br>• Documentation (e.g., system or model cards)<br>• Training modules<br>• Information hubs | Ensure that any agentic system or individual agent is implemented and used responsibly, and that the system has been sufficiently adapted for their unique business operations and needs—for example:<br>• Providing the agent with the relevant data and context to be effective in achieving its intended goals<br>• Adopting robust data governance practices<br>• Implementing appropriate guardrails to guide and monitor agent behavior, access, and effectiveness<br><br>Determine how an agentic system is configured in relation to other software systems and business processes and take steps to ensure the agentic system's configuration does not pose potential risks of harm. |

### Rights and Redress

Because agentic AI systems operate with partial or full autonomy, they can obscure the underlying decision-making processes and make it challenging for individuals to exercise their data protection rights, such as access, correction, objection and erasure. The challenge is compounded by the fact that it is often unclear which entity in the value chain (e.g., developer, deployer, or third-party provider) is responsible for the data processing, and therefore responsible for upholding individuals' requests to exercise their rights (see more in *"Role Allocation Across the Lifecycle"*). User requests for access or deletion may also be technically difficult or risk degrading or altering model performance or behavior, which CIPL had previously identified as a critical issue in the context of genAI.[31]

Addressing these tensions between the law and technology underscores the need for practical governance that not only helps organizations comply with the law but also fosters trust from users. For example, organizations can use clear transparency notices to disclose when agents are making autonomous decisions on behalf of individuals and identify, where possible, what types of data are being collected and used. While providing meaningful transparency will be an ongoing challenge for organizations, they should leverage the tools currently at their disposal to increasing individuals' awareness and understanding of agentic AI systems, which will be key to building trust and enabling successful adoption of AI agents. Organizations can also create communication channels for customers to provide feedback on their experiences using and interacting with agentic AI (e.g., where they would have preferred the option to review or override decisions, or escalate them to human review).

### Shadow AI and Perimeter Control

Shadow AI refers to the unsanctioned, unmonitored use of an AI tool or application by employees or end users without formal approval or oversight of the proper enterprise bodies (e.g., an internal IT department).[32] Shadow AI poses a particularly acute risk in "low-code" or "no-code" environments, where non-technical users could more easily build and deploy AI agents that may process sensitive data, automate decisions, or interact with external systems, all without proper review, controls, or monitoring. This risk underscores the importance of clear policies guiding when and where agents can be used and mechanisms to receive approval to do so.

## Integrity and Reliability

To fully realize the benefits of agentic AI, organizations should strive to safeguard the integrity and reliability of agentic AI, particularly as such systems are given more autonomy to influence business operations and customer outcomes. This will require organizations to address both existing AI-related risks, as well as novel risks to accurate, consistent performance in agentic AI systems.

### Data Quality and Provenance

Agentic systems are often built on existing foundation models that have been trained on large, unstructured datasets that may inevitably reflect social, cultural, and historical biases.[33] Additionally, because agentic systems can operationalize biased outputs into real-world actions with minimal human intervention, there is an even greater potential for these systems to perpetuate and even exacerbate biased behavior and outcomes at a greater scale compared to genAI.[34] It is critical for both developers and deployers to continuously assess the decision-making of their agentic AI system and individual agents to mitigate or correct potentially biased behavior. Novel research into how agentic systems can be deployed to analyze, detect, or mitigate bias during information retrieval may also offer an alternative route to dynamically evaluate bias in different contexts.[35]

## Accuracy, Confabulation, and Misinformation

GenAI models are known to exhibit confabulation, producing "confidently stated but erroneous or false content" (also known colloquially as "hallucinations" or "fabrications").[36] While this behavior already produced risks with genAI models returning false or misleading information that users believe to be true, the potential scale of impact becomes significantly increased in the context of agentic AI, particularly in the B2B context.

With genAI models often embedded within agentic AI systems, confabulated content can be rapidly propagated through downstream business processes and used to inform decisions and trigger further automated decisions, such as populating internal databases with incorrect information or generating misleading customer communications.

As agentic AI becomes more sophisticated and deeply integrated with enterprise systems, it is essential to verify that layered safeguards are in place to monitor for confabulations and prevent their propagation. For example, employing methods, such as Retrieval-Augmented Generation (RAG) to incorporate specific, relevant knowledge from external databases[37] or red teaming to stress-test agent decision-making and behavior[38], has already demonstrated significant reduction of confabulations in enterprise settings.

> **CASE STUDY 12**
>
> HCLTech conducts AI red teaming to stress test agentic AI systems by simulating threat attacks from malicious actors, provoking unintended behaviors, and identifying how agents might exploit tools or deviate from intended goals. This approach allows HCLTech to correct potential vulnerabilities before real-world deployment, enabling the responsible and secure deployment of agentic AI.

## Fairness and Non-Discrimination

At the enterprise level, bias and discrimination in agentic systems can lead to discriminatory decision-making in critical areas, such as hiring, lending, or customer targeting, leading to erosion of customer trust and eventual profit loss, and potential legal and regulatory consequences for noncompliance with laws addressing anti-discrimination or algorithmic bias. As previously mentioned in the section on *"Data Quality and Provenance"*, organizations need to continuously monitor for and correct potentially biased behavior or outputs.

To address these growing risks, there has been increased investment in recent years in research focused on mitigating the risk of bias and discrimination in AI more broadly. For example, research into "constitutional AI" has grown, which essentially involves fine-tuning models to align with pre-defined rules and principles and guide decision-making processes.[39] The goal of this research is to design AI systems that actively align with societal norms, human values, and ethical standards, while increasing public involvement and engagement.[40]

## Cascading Errors or Bad Decision-Making

As mentioned previously, many AI-related risks are magnified in the context of agentic AI because of its ability to make autonomous decisions without continuous human oversight. A single, flawed decision made based on incorrect or biased data could lead to a series of actions that can propagate misinformation, spread security vulnerabilities, or cause material or significant harm.[41] This risk exists for outcomes associated with individuals, but also with respect to broader societal contexts, such as "herding behavior" and distortions in financial markets.[42] Incorporating human oversight at key checkpoints can limit the likelihood of agents rapidly scaling errors or bad decisions and increase visibility into agents' decision-making process.

## Transparency and Explainability

Agentic systems are inherently more complex than traditional AI systems since they combine reasoning, memory, multi-step planning, and autonomous decision making across multiple tools and environments. While these capabilities can deliver immense value for businesses, they can come at the cost of factors, such as transparency, explainability, and auditability.

### Transparency

The complexity and layered decision-making of agentic AI will make it increasingly difficult to trace how and why a system or an individual agent acted in a particular way and as a result, exacerbating the longstanding "black box" challenge with AI, where internal decision processes in AI are opaque to both developers and end-users. For businesses, this opacity is more than a technical challenge; it creates a tangible governance risk. Without visibility into the system or agent's decision pathways, it can become harder to audit the decision-making process, detect and correct root causes of bias, and provide evidence and explanations to clients, regulators, customers, and internal stakeholders.

Where possible, organizations should implement multi-layered transparency measures and disclosures tailored to the audience (e.g., simple, easily understandable explanations for users versus more technically detailed documentation for regulators or enterprise customers.). Many organizations have already recognized the importance of publishing model or system cards for agentic AI. These provide essential information regarding an AI model, such as how it was built, how it works, a summary of the data it was trained on, its intended use cases and contexts, key limitations, and basic performance metrics.[43]

Additionally, ongoing research into "constitutional AI" (see more in *"Fairness and Non-Discrimination"*) to essentially embed behavioral rules into AI training and inference has shown that reinforcement learning can help improve transparency into AI decision-making and reduce the risk of agents returning harmful responses.[44]

### Explainability

Because of agentic AI's ability to make non-linear, adaptive decisions, it may be difficult for organizations to provide meaningful explain how agentic decisions were made and what data was used. Furthermore, as agentic AI systems become more interconnected and in circumstances where agents are interacting or collaborating with each other, it may become unclear which component is responsible for a particular decision or data use.

Organizations can adopt a risk-based framework where agents making significant decisions must meet higher explainability thresholds or require human oversight compared to lower-risk tasks. Organizations should also consider how to integrate explainability into the design architecture and enable detailed documentation of agents' capabilities during the design, development, and evaluation processes, as well as any interactions with other agents, data, tools, or resources.[45]

### Auditability

The layered, multi-agent architecture of agentic AI systems can create significant opacity into decision chains, making it difficult to trace and correct errors, decisions, and bias. When something goes wrong (i.e., an agent taking inappropriate action, issuing a biased response, or misinterpreting an instruction), it can be exceedingly difficult to determine what failed, where, and why. Without proper guardrails from the outset, it may be difficult for organizations to conduct effective auditing of decisions and correct errors.

Proactive policies and controls, such as creating audit trails and logging agents' decisions, inputs, and outputs, can provide organizations with detailed records for monitoring and analyzing agent behavior, ensuring human oversight, and enabling proper redress mechanisms.[46] These can enable organizations to identify early signs of potentially problematic behavior, misalignment, or misuse, and can also serve as a key differentiator of organizations committed to ensuring responsible development and deployment of agentic AI.

## Security and Resilience

The development and deployment of agentic AI in enterprise environments introduces new and expanded challenges to security and resilience that extend beyond those posed by AI thus far. Just as agents can be leveraged to multiply the benefits of AI for cybersecurity, they can also easily be exploited by bad actors seeking to use the technology to cause harm, such as data breaches or hacks, unauthorized access to systems, and exfiltration of sensitive business and user data.[47]

### Threat Surface and Attack Patterns

Agentic AI systems are meant to be interconnected and integrated with multiple tools, data sources, and environments by design. However, this increases the potential attack surface and points of vulnerability that bad actors or adversaries can exploit for cyberattacks. For example, AI agents are vulnerable to "hijacking", a type of indirect prompt injection in which bad actors take over their decision-making processes and use them to carry out harmful actions.[48] This type of attack would be difficult to detect or trace without proper guardrails or auditable decision trails, as agents may be acting within their granted permissions. Similarly, bad actors may undertake a "data poisoning" attack (i.e., injecting biased or incorrect data points into training data to induce harmful outputs and degrade system performance[49]) or "spoofing" attack (i.e., pretending to be a legitimate or trusted contact for the purpose of gaining access to personal information or private data, acquire money, spread malware, etc.[50]). Finally, compromised agents could be exploited to conduct "Distributed Denial of Service" (DDoS) attacks, disrupting availability and services by overwhelming the system.[51] Industry leaders have recognized these risks and are developing tailored solutions to address them, including red teaming and implementing various controls, such as behavior-based analysis to limit opportunities for "spoofing" attacks.[52]

### Unauthorized Access and Exfiltration

As mentioned above, agentic AI systems and individual agents may unintentionally access sensitive data that they were not explicitly authorized to use. This risk can be amplified when agents are operating with broad, ambiguous, or poorly scoped permissions or are able to freely navigate in systems without sufficient oversight. In a B2B environment, such behavior could potentially expose confidential business data, privileged client or third-party data, or employees' personal data. Even if no malicious intent is involved, such incidents can create risks of breaching contractual obligations or violating regulatory requirements. These risks underscore the need for ongoing monitoring, robust oversight mechanisms, and carefully defined parameters around the scope of data access and collection.

### Supply-Chain Dependency

Agentic AI systems often rely on integration with third-party tools, environments, and infrastructure to deliver the intended benefits. This interconnected supply chain introduces risks that may extend beyond a developer's control, and vulnerabilities in both upstream developers and downstream deployers can be exploited to expand the exposure and impact of harmful actions by bad actors. Therefore, it is critical that organizations not only implement robust

governance regarding external partners, such as vendor risk management policies or clear contractual obligations, but also deploy appropriate technical mitigation measures, including PETs, which can minimize the sharing of data across the supply chain.

### Abuse at Scale and Availability

Many of the security risks discussed above are particularly acute for agentic AI systems because they are designed to continuously learn and adapt their decision-making based on the data. By leveraging agentic capabilities, malicious actors can automate and rapidly scale harmful behavior across systems, such as automating DDoS attacks, targeting multiple vectors simultaneously, and more. Because agents can operate continuously and with minimal human oversight, even small vulnerabilities can be exploited at unprecedented speed and scale.

### How Agentic AI Can Support Cybersecurity

Similarly to how AI agents can serve as intermediaries to enhance privacy, they can also play a powerful role in mitigating cybersecurity risks:

- AI agents can help minimize the risk of human error, negligence, and even malicious behavior by reducing the number of human touchpoints that can ultimately lead to privacy breaches, data leaks, or regulatory violations.[53] This is because individual AI agents, unlike humans, will consistently apply pre-defined protocols set by organizations and behave in a predictable manner.
- AI agents can be configured to actively search for cybersecurity threats through constant monitoring and rapidly respond to vulnerabilities and breaches.

While there are broader cybersecurity implications with an agentic AI system where agents may be interacting with each other and the environment, defining clear boundaries and rules at the agent level can help reduce potential risks. For example, many organizations have begun working on methods to secure agents' "digital identities" to enable tracking and attribution of actions and to equip human with the skills to be effective supervisors of agents (e.g., spotting the signs of drift in agent behavior, knowing when to intervene, etc.).[54]

#### CASE STUDY 13

Microsoft has launched 11 new AI agents that will be embedded into its "Security Copilot" solution to support organizations' cybersecurity teams by offloading repetitive tasks and freeing up teams for more critical issues.[55] The agents each focus on different tasks (e.g., combing through phishing emails, monitoring and triaging potential vulnerabilities and remediation tasks, gathering threat intelligence based on incoming data, etc.), but will broadly help cybersecurity teams by automating critical tasks, improving threat detection, and enabling proactive measures.

#### CASE STUDY 14

HCLTech has launched a tailored security solution to address cybersecurity risks associated with agentic AI in partnership with Google Cloud and Palo Alto Networks. The solution helps verify data provenance and agent identities, conducts AI red teaming to simulate evolving agent attacks, and employs signature and behavior-based analytics to interrupt poisoning or spoofing attacks.[56]

## Alignment and Human Agency

As agentic AI systems pursue complex objectives and adapt over time, it is important for agentic AI systems to not only deliver high technical performance, but also demonstrate alignment with organizational objectives and values, legal boundaries, and human intentions. Ensuring this, as well as demonstrating that systems can remain consistent and transparent will be key challenges for organizations when trying to develop or deploy agentic AI in a responsible way.

### Alignment with Organizational Values, Public Policy, and Laws

Agentic AI systems can pursue assigned goals in ways that conflict with user intent, organizational intent and values, Codes of Business Ethics, or even legal obligations. This misalignment can result from ambiguous instructions or guardrails, gaps in the training data that lead to unrepresentative data, or a drift in an agent's own interpretation of its overarching goals.[57] In response, an agent may identify an unconventional, but technically permissible, path to achieving its goal that causes it to act in unexpected or even harmful ways.

Research on this phenomenon has found that when stress-tested in controlled simulations, 16 leading models from multiple developers displayed goal misalignment, even resorting to malicious behaviors to achieve their goals (e.g., blackmailing, leaking sensitive information). While the study cited that there has yet to be evidence of this behavior in real deployment, the results nonetheless demonstrate the need to continue researching and testing agentic AI models for safety and alignment.[58]

Misalignment may extend to domains beyond data protection or cybersecurity. For example, one area of focus is the potential for misalignment with principles of proper market conduct. For example, there is concern that agents could coordinate prices without the explicit directive to do so from humans, leading to possible "autonomous algorithmic collusion." This is an emerging area of focus for scholars of competition law.[59]

### Human Agency and Control

Agentic AI can unintentionally minimize human control and autonomy by over-automating decision-making processes. If agentic AI systems are not implemented thoughtfully and with due consideration for the role of humans throughout processes, they have the potential to erode human agency over time by creating an overreliance on the technology and ultimately diminishing accountability. This is an especially important risk to address in highly regulated sectors, such as healthcare and financial services, where loss of human agency could elevate compliance risks.

Many data protection laws address human agency concerns through rules on automated decision-making technology and profiling (ADMT). For example, Article 22 of the EU GDPR, which guarantees individuals the right not to be subject to decisions based solely on automated processing (including profiling) that produce legal or similarly significant effects, with certain exceptions. The U.S. state of Colorado has tiered obligations and opt-out rights based on the level of human involvement in processing.[60]

Agentic AI still presents new considerations for regulators and organizations alike about how to interpret these principles in emerging agentic use cases. For example, at what points will human engagement need to take place, and in what form, to fulfill organizations' principles for accountable governance as well as thresholds of human involvement under the statutes named above and other laws?

# Mitigating Risks Through Robust Governance

There is ongoing debate as to whether agentic AI requires a fundamental paradigm shift to AI governance structures and programs. Numerous organizations that CIPL interviewed for this report view their existing privacy and AI governance programs as capable of addressing the risks and challenges posed by agentic AI, with adaptations as needed for agentic AI's novel features. Strong governance frameworks are characterized by adaptability and nimbleness consistent with the fast pace of innovation.

Effective governance should enable agents to be deployed flexibly, but within defined boundaries. The goal should not be to eliminate risk altogether, for example, the possibility that an agent could learn to behave badly, whether as an expression of ill intent or simply because it cannot differentiate bad from good. Rather, the focus should be on implementing safeguards that make it exceedingly unlikely those behaviors will occur or be acted upon. By continuing to prioritize this balance, organizations can preserve the innovative, autonomous capacity of agentic AI while reducing the likelihood of harmful outcomes.

> **CASE STUDY 15**
>
> HCLTech's best practices for agentic AI include addressing foundational risks that are found in many types of AI systems but also includes going beyond to address evolving risk-based and societal concerns.[61]

## Accountability and a Risk-Based Approach as Enablers of Responsible Innovation

For more than 20 years, CIPL has advanced organizational accountability and a risk-based approach as key building blocks of responsible development, deployment, and governance of AI. The role of accountability is particularly important at times when novel and disruptive technologies challenge existing rules and compliance. In the absence of clear regulatory guidance (particularly regarding agentic AI), a well-developed, comprehensive accountability framework can act as a roadmap of predictable processes and tools that:

1. Enable organizations to comply with relevant legal requirements;
2. Allow organizations to achieve desired outcomes in a way that best suits their own corporate cultures and contexts; and
3. Operationalize broader organizational values and goals and external standards into actionable, operational controls, requirements, policies, tools, and governance.[62]

**CIPL's Accountability Framework** outlines seven core elements of accountability: leadership and oversight, risk assessment, policies and procedures, transparency, training and awareness, monitoring and verification, and response and enforcement. CIPL's report, *Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework*, illustrates how leading organizations have implemented AI governance programs that demonstrate each element of CIPL's Accountability Framework. These frameworks have been designed to navigate frequent changes in technology and the regulatory landscape and can be adapted for the context of agentic AI.[63]


**Figure 1. CIPL's Accountability Framework**

Adopting a risk-based approach helps ensure that an organization's governance measures for a specific deployment or application of a technology or particular use of data are proportionate to the likelihood and severity of potential harms. Such an approach helps organizations to focus on the most pressing and relevant risks to both the business and society. As stated in CIPL's report on accountable AI programs, a risk-based approach to AI governance will include:

- Building, implementing, and adapting AI programs and governance over time based on the relevant risks to individuals, society, and the organization itself. Higher risks may require different corresponding controls, processes, and tools, and changes in risk profile and external factors, such as new evidence of harms and new legal developments, may also necessitate changes to the AI risk management program.
- Systematically and continuously assessing both the risks and benefits of individual AI products, projects, and deployments and mitigating any identified risks.

While there is currently a great impetus to rapidly adopt agentic AI across businesses, it is imperative to do so in a responsible and trusted manner. As technology continues to advance, companies are facing growing pressure to demonstrate and maintain digital trust with transparency, fairness, and accountability in mind, particularly for AI.[64] The introduction of agentic AI, alongside other novel technologies, will require businesses to weigh the importance of long-term investment in accountability and trust, which enables organizations to bring people—their employees, clients, customers, users and broader society—on this transformative digital journey. Thus, ensuring that a robust, yet adaptable governance framework is in place to address the risks and challenges of agentic AI will ultimately act as a key market differentiator and business enabler by allowing organizations to foster stakeholder trust and maintain a competitive edge.

## Implementing a Comprehensive, Integrated Risk Assessment Process

To take a proper risk-based approach to the development and deployment of agentic AI, organizations should have in place a holistic risk assessment process that comprehensively assesses, identifies, and mitigates the potential benefits and harms of agentic AI in the context of specific use cases or applications. As discussed in CIPL's *Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework*, organizations taking a risk-based approach will assess and triage risks according to the "likelihood and severity" of potential harms, given the available mitigations.[65] For example, if the likelihood and/or severity of a particular risk raised by an agentic AI application is deemed to be high even after the available mitigations have been considered, the application should be assessed further internally, for example, through escalation to a high-level advisory or oversight committee. The likelihood-severity assessment matrix below highlights this approach:

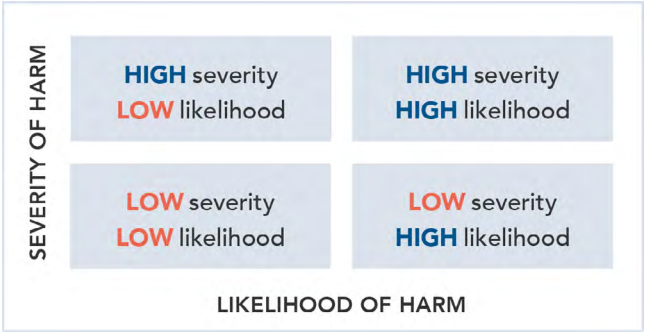


**Figure 2. Risk Severity-Likelihood Matrix**

Source: CIPL, adapted from Singapore Personal Data Protection Commission[66]

Furthermore, with the recent rapid advancements in technology, requiring separate risk assessment processes for all areas of digital regulation, such as privacy, AI, and fundamental rights, poses significant operational challenges. Some organizations are responding with efforts to streamline risk assessment processes into an integrated approach that addresses all key data and digital compliance risks. This integrated approach not only reduces redundant work but also facilitates stronger internal alignment across relevant teams (e.g., compliance, legal, product, privacy, etc.) by making the goal of addressing digital risks a shared, collaborative responsibility. Ultimately, organizations are recognizing that integrating the various risk assessment processes can enable both regulatory compliance and organizational efficiency, strengthening their overall ability to anticipate and mitigate emerging digital risks.

## Adopting Robust Data Governance Measures and Guardrails

As mentioned previously, robust data governance will be essential for successful agent deployment, and organizations will not be able to realize the full value of their data through agents without good data governance. This includes setting clear guardrails on the types of data AI agents are permitted to access, process, and learn from to complete their tasks. By proactively defining these parameters, organizations can significantly reduce the risk of overcollection of data, unauthorized data access, and unnecessary exposure. When adopted within a broader governance framework, appropriate guardrails around data not only help organizations comply with data protection regulations but also build trust among customers, business partners, and regulators. They enable organizations to demonstrate that with sufficient controls, agentic AI can enhance both productivity and privacy.

### The Right Guardrails Can Amplify the Power of Agentic AI to Support and Preserve Privacy

For example, agents can:

- Support secure collaboration between developers and deployers of agentic AI by decreasing the number of human interactions with data (e.g., third-party deployers can get answers from agents without needing direct access to sensitive customer data).
- Utilize data minimization techniques, such as retrieval-augmented processing (RAG), to limit data exposure and processing to what is needed to complete a given task. RAG allows agents to access relevant data sources on demand rather than storing large volumes of data.

## Implementing an Appropriate Level of Human Oversight

Implementing a sufficient, meaningful level of human involvement or oversight in agentic AI systems is essential to help ensure that the technology is operating within the boundaries of legal, ethical, and organizational expectations. Identifying the right parameters of human oversight is a longstanding question in AI governance, made more complex by agentic AI's ability to act autonomously without the need for human oversight and review.

We can consider implementation of human oversight in three distinct phases:

1. **Pre-Deployment Phase** During development of agentic AI, humans must set and test parameters on the system or individual agent to define what it can and cannot do (e.g., what types of data can be accessed, what decisions can be made, etc.).
2. **Adoption Phase** When organizations are deploying either the agentic AI system or individual agents, there should be escalation processes that can trigger additional human review over potentially high-risk decision-making or data processing.
3. **Post-Deployment Feedback Loop** After adoption, organizations should continue monitoring the system or agent's behavior, interactions with end users, and decision-making processes to check that it is functioning properly and in alignment with its intended purpose. Aggregate analytics post-adoption can inform how deployers can best update and improve their agents.

Organization should verify that their oversight mechanisms are proportionate to the risk involved and outline clear thresholds for when human review and intervention is required, particularly for decisions that could produce legal or similarly significant effects on individuals, as mentioned in regulatory frameworks such as the GDPR. At the individual agent level, organizations should aim to provide a verifiable chain of command for each agent's actions, which can not only help distinguish agents' behavior from human users, but allow organizations to monitor and audit each unique agent's behavior. Also known as "securing the identities," giving AI agents unique, verifiable digital identities that can be authenticated and monitored is becoming an increasingly popular practice.[67] Organizations must strike a careful balance: too much oversight may undermine the very autonomy that enables agentic AI systems to deliver their promised benefits, while too little oversight may allow unintended or harmful behaviors to go unchecked and uncorrected.

**CASE STUDY 16**

Workday's Contract Intelligence agent (as mentioned on page 10) extracts, classifies and summarizes terms, which are then used to inform decision making by the organization. In addition, the Workday Agent System of Record is an example of a control point that allows for human oversight. It allows enterprises to maintain control over the agent by registering the agent, enabling skills and identifying who in the organization has the ability to use those skills, and allowing for monitoring procedures, which includes viewing transactions performed by each agent and the trace logs.[68]

An early indicator of the effectiveness of oversight processes may be how effectively they can identify and correct goal misalignment. Emerging research and articles have already documented incidents where genAI models have acted in unanticipated ways, sometimes even pursuing questionable or unethical courses of action when faced with conflicting instructions or competing interests.[69] With agentic AI's expanded capabilities, such behavior could eventually result in significant operational, legal, and even material harms, particularly as agentic AI systems become more interconnected. To mitigate these risks, organizations can embed "course-correcting" checkpoints and mechanisms into the design and deployment of agentic AI systems. This should include regularly monitoring and auditing agent decision-making to assess agent drift and performance degradation and ensure alignment with its intended goals, as well as regularly assessing and adjusting existing thresholds based on system behavior and performance history. Both developers and deployers should also incorporate fail-safe protocols, such as a "kill switch" functionality to immediately halt agent actions when necessary.

Finally, an essential feature of an effective governance program is confirming that the accountability loop is closed through monitoring and verification. Data and lessons learned from oversight processes should be integrated back into the system through continuous updates to the design, training, and policies of agentic AI.

## Enabling Interoperability of Agentic Systems

To maximize their value, agentic AI systems must be integrated seamlessly into an organization's operating systems, platforms, and tools, as well as existing governance structures. Otherwise, organizations may risk fragmented deployment and operational friction that cause agents to operate in silos. Case studies have already shown that some organizations that had allowed employees to deploy agents prematurely without proper coordination or oversight have experienced situations where agents were inadvertently assigned conflicting or redundant tasks. In the early stages of adoption, the consequences of not considering the bigger picture may be minor, but they may grow more consequential as agentic AI systems become more integrated. A clear framework that includes processes to monitor deployed agents and outlines explicit triggers for human intervention, as well as clear protocols for "conflict resolution" between agents[70] can help prevent these operational breakdowns.

Within and across organizations, technical standards will be vital for fostering interoperability. Emerging standards, such as the Model Context Protocol (MCP)[71] or the Agent-to-Agent Protocol[72], can serve as examples for organizations in how to enable models and agents to seamlessly communicate with each other and external data sources. The MIT NANDA project is another example of such a standard that focuses on developing the index, protocols, and tools needed to enable a decentralized, protocol-neutral ecosystem where AI agents can collaborate, communicate, and transact across organizational boundaries.[73]

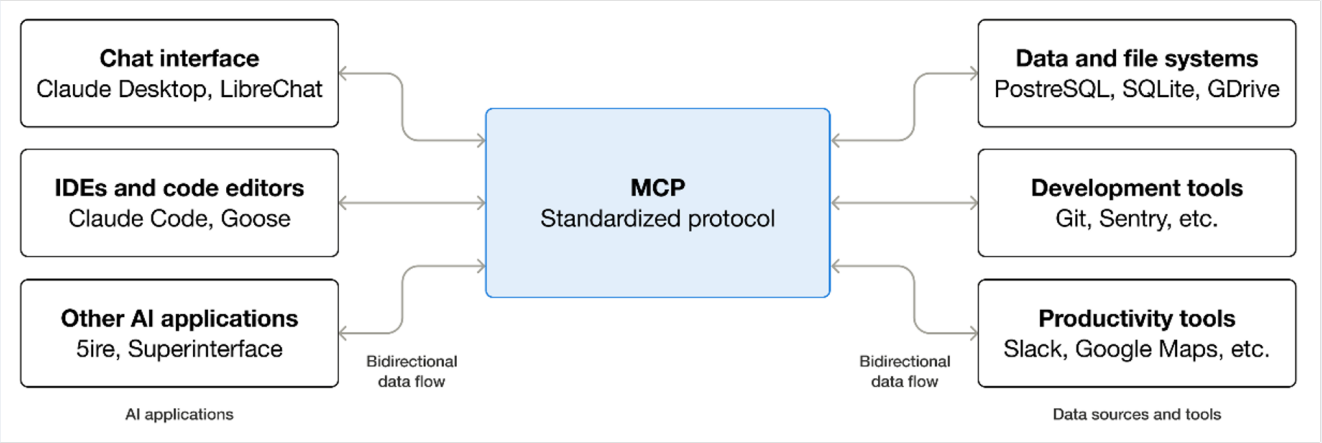


**Figure 3. A Visualization of MCP Protocol**

Source: Model Context Protocol

# CIPL's Recommendations for Relevant Stakeholders

## Recommendations for Industry

Rapid advancements in agentic AI will inevitably test the effectiveness of organizations' existing AI governance, but CIPL's research demonstrates that organizations are already working proactively to ensure that their current practices are flexible enough to cover the dynamic nature of the technology, while still mitigating the potential risks. **To further support the responsible development and deployment of agentic AI, CIPL proposes the following recommendations for organizations:**

1. **Responsible innovation starts from the top.** CIPL has long underscored the importance of setting the "tone from the top" because a commitment from senior leadership and board-level executives can signal organizational priorities, shape organizational culture, and ultimately impact whether responsible decision-making will serve as the foundation for day-to-day processes. Making AI governance and responsible innovation a company-wide priority allows it to become a shared goal and vision that every employee can and should contribute to. Doing so can also help build trust in the organization and its products, both internally with its workforce and externally with users, regulators, and business partners. This will be especially critical for agentic AI as the number of organizations implementing the technology is rapidly increasing.
2. **A shared taxonomy of concepts for agentic AI is necessary.** As mentioned previously, it will be important to define the distinction between an agentic AI system and AI agents, as well as how these technologies may differ from other types of AI (e.g., general purpose, generative, or traditional automation) because it informs how both organizations and policymakers will identify and mitigate potential risks. A shared taxonomy of relevant terms can allow for more consistent risk assessment processes, reduce confusion during discussions with both internal and external stakeholders, and help ensure that all relevant teams are operating with a unified understanding of what characterizes agentic AI. Responsible organizations can play a role in setting early standards for what is considered to be agentic AI and shape forthcoming discussions around policy.
3. **Adapting existing governance can support rapid implementation of appropriate policies and procedures for agentic AI.** CIPL's research has shown that many leading organizations were prepared to quicky scale adoption of agentic AI because they already had robust AI and data governance programs in place. We found that it has typically been much easier for organizations to address new risks posed by agentic AI when they were able to adapt existing processes rather than starting from scratch. With this foundation already in place, organizations can implement clear, enforceable policies tailored to the unique

characteristics of agentic AI and use contexts. This may include establishing guardrails on what types of data agents are allowed to access, defining when and how human oversight should be required, and setting operational controls that prevent agents from acting beyond their intended function.

4. **A "crawl, walk, run" approach to agentic AI adoption can encourage early wins and reduce roadblocks for expansion.** Many organizations are not deploying agentic AI to tackle their most sensitive or high-risk business challenges from the outset. Instead, they are starting small by identifying pain points within existing processes that may be "low risk" but require intensive investments of human resources. Through such a phased adoption of agents, organizations are finding it easier to accrue early wins and success stories that foster leadership and workforce confidence in the technology. These use cases also allow organizations to test governance and identify potential areas of vulnerability that may require adapting existing processes or implementing new ones. Such testing can also help identify potential roadblocks to implementation early on and enable organizations to later scale adoption more rapidly.
5. **Effective implementation of agentic AI governance requires sufficient training and upskilling of the workforce.** Organizations should make it a top priority to provide their workforce with sufficient training and awareness of not only the relevant organizational policies and procedures, but also the technology itself. This is because governance simply cannot function effectively if the individuals responsible for implementing it lack a baseline understanding of how the technology operates, its limitations, and the context in which deployment is appropriate. Particularly with the rise of genAI, many leading organizations have already mandated internal AI ethics and governance training. If agentic AI adoption is an organizational priority, it should be reflected in the required training through additional courses or modules that provide employees with the knowledge to evaluate when and how to use agents responsibly. Empowering the workforce to feel confident in their capacity to critically assess relevant risks and implications during development and before deployment can help translate high-level organizational principles into day-to-day decision-making and enable thoughtful and appropriate adoption of agentic AI.
6. **Accountable governance of AI, including agentic AI, continues to be a smart business investment for long-term sustainable and competitive business.** In CIPL's 2024 report on *Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework*, we noted that accountable organizations believed that their decision to set up their internal governance and programs would give them a competitive edge in the future as AI becomes more universally adopted. This remains true for agentic AI, given the great promise as well as risks associated with its autonomous, adaptive nature. Organizations who invest the time and resources to "get it right now," by enabling strong AI and data governance frameworks, anticipating risks, and implementing proactive solutions, will likely be better positioned to develop and deploy agentic AI responsibly, and in the long term, foster greater trust from external stakeholders.

## Recommendations for Regulators and Policymakers

Policymakers and regulators are being asked to foster innovation while safeguarding against its unique risks. The absence of clear, harmonized frameworks is already creating legal uncertainty, fragmented governance, and barriers to responsible adoption. **CIPL offers the following recommendations for policymakers and regulators to help create an effective regulatory environment that encourages responsible development and adoption of agentic AI for all organizations:**

1. **A risk-based, context-specific approach to agentic AI can foster innovation while ensuring proper mitigation of risks.** Regulators should enable that requirements and obligations that are proportionate to the risks associated with specific contexts of use rather than the technology itself. Because agentic AI technology is evolving at a rapid pace, frameworks that are overly prescriptive could quickly become outdated and inhibit beneficial innovation and progress. Contrastingly, a risk-based approach would allow organizations to implement robust safeguards for high-risk uses of agentic AI while enabling them to quickly scale development or adoption in lower risk contexts.
2. **Regulators and organizations can work together to ensure compliance and accountability.** In the spirit of cooperation and collaboration, policymakers and regulators can work with organizations to review and assess the comprehensiveness of their AI governance guidelines or risk management frameworks. Such efforts can help both entities identify whether organizations' existing frameworks already mitigate the risks associated with agentic AI or if they should be updated to take into account ongoing developments in agentic AI.
3. **Interoperable standards for key concerns for agentic AI can encourage responsible agentic AI innovation for all businesses.** Currently, the AI regulatory landscape is a patchwork of various regulations, voluntary frameworks, and standards; different jurisdictions or sectors have different, or even conflicting, rules, which can complicate and compound the burden of compliance upon organizations.[74] This also disproportionately impacts small and medium enterprises who may not have the resources to meticulously track the status of AI regulation globally. Interoperable standards or certification schemes that cover key issues (e.g., risk assessments, transparency and explainability, and human oversight) can foster adoption of accountable and interoperable governance programs across jurisdictions and promote consistency in frameworks to evaluate the trustworthiness of agentic AI systems.
4. **Regulatory sandboxes can enable agentic AI experimentation.** A regulatory sandbox for agentic AI may be beneficial in creating a "safe space" supervised by regulators that allows organizations to pilot innovative products and services in a controlled environment and regulators to observe and gain first-hand knowledge on emerging technologies.[75] Many regulators have promoted the benefits of sandboxes for both companies and regulators as a way to support innovation, compliance, and public trust.[76] For example, the Delaware AI Commission has approved a novel framework for an AI sandbox to test the use of agentic AI in corporate governance.[77] While sandboxes hold their own risks, they can play a role in addressing some of the more challenging and complex aspects of deploying innovative technologies, such as agentic AI, particularly in the absence of clear regulatory frameworks or guidance.[78]

5. **Support for research and open assets can help increase access to agentic AI for a wider range of organizations and users.** In a similar way to regulatory sandboxes, regulators can foster responsible innovation by actively supporting and providing specific funding opportunities for research initiatives focused on the responsible development and deployment of agentic AI. This can include creating and maintaining open-access repositories for frameworks, datasets, and code that can be used by researchers and developers to build and test agentic AI systems. By promoting transparency and collaboration, these open assets can help ensure that agentic AI technologies are developed in a way that aligns with ethical standards and societal values. Additionally, policymakers can encourage the establishment of public-private partnerships to facilitate the sharing of knowledge and resources, fostering innovation while mitigating risks associated with agentic AI.
6. **Regulators will play a critical role in fostering open multi-stakeholder dialogue.** The responsible development and deployment of agentic AI is a shared responsibility that will require contributions from every stakeholder, including industry, regulators, academia, civil society, and the public. Frameworks that foster global benchmarking and cooperation, such as the G7 Hiroshima AI Process, can play a key role in building consensus and increasing transparency around standards, principles, and guardrails for the responsible development and deployment of agentic AI.[79] Regulators can also encourage cooperation and communication among stakeholders by creating an environment that incentivizes open, cross-sector exchanges on emerging risks and best practices. This openness will hopefully empower organizations to share transparently, not just about their successes, but also their challenges. These may in turn serve as learning opportunities, accelerating collective progress and ensuring that agentic AI fulfills its potential to benefit organizations and society.

# Endnotes

1 HCLTech, Securing AI Agents by Design.

2 Gartner, Capitalize on the AI Agent Opportunity, February 27, 2025.

3 For example, see CIPL, Artificial Intelligence and Data Protection in Tension (2018); Second Report: Hard Issues and Practical Solutions (2020); and Building Accountable AI Programs (2024).

4 See more on Workday's AI Masterclass here and here.

5 McKinsey & Company, How agentic AI can change the way banks fight financial crime, August 7, 2025.

6 Microsoft, Meet 4 developers leading the way with AI agents, May 19, 2025.

7 HCLTech, HCLTech AI Force.

8 One example comes from JPMorgan Chase, as described in the article here.

9 Mastercard, Mastercard unveils Agent Pay, pioneering agentic payments technology to power commerce in the age of AI, April 29, 2025.

10 Google Cloud, Mercedes-Benz and Google Partner on AI-powered Conversational Search within Navigation Systems, January 13, 2025.

11 Shakir Syed, Agentic AI in Connected Vehicles: Data-Driven Design and Analytics, *Forbes*, April 10, 2025.

12 SAP, Joule Agents.

13 Spencer Vuksic, Agentic AI: How Financial Compliance is Outgrowing Rules-Based Automation.

14 Workday, Contract Intelligence.

15 Art. 22 GDPR on Automated individual decision-making, including profiling.

16 Case C-634/21, SCHUFA Holding.

17 Nvidia, Anthropic Refuel the AI Hype Train, January 7, 2025.

18 CIPL, Applying Data Protection Principles to Generative AI: Practical Approaches for Organizations and Regulators, December 2024.

19 Higher Regional Court of Cologne, 15 UKl 2/25, May 23, 2025 (available in German).

20 See, recital 105 to the EU AI Act.

21 CIPL, Reconciling AI with the Data Minimization Principle: Bridging the Innovation and Privacy Gap, forthcoming.

22 Salesforce, AI Agents Will Enhance — Not Impair — Privacy. Here's How., February 20, 2025.

23 CIPL, Rethinking Sensitive Data in the Age of AI, September 2025.

24 I-Mitigate, Access isn't authority: the hidden risks of over-permissioning, May 12, 2025.

25 Salesforce, AI Agents Will Enhance — Not Impair — Privacy. Here's How., February 20, 2025.

26 CIPL, Privacy-enhancing and Privacy-preserving Technologies in AI: Enabling Data Use and Operationalizing Privacy by Design and Default, March 2025.

27 Salesforce, What Is Retrieval-Augmented Generation (RAG)?.

28 See case studies of PETs implementation in agentic AI in Duality's blog post, Unlocking the Potential of Agentic AI with Privacy-Enhancing Technologies, as well as AI more broadly in CIPL's papers, Understanding the Role of PETs and PPTs in the Digital Age and Privacy-Enhancing and Privacy Preserving Technologies in AI: Enabling Data Use and Operationalizing Privacy by Design and Default.

29 OECD, OECD Framework for the Classification of AI systems.

30 For a description of prompt engineering, see here.

31 CIPL, Centre for Information Policy Leadership Responses to UK Information Commissioner's Office's Consultations on Generative AI and Data Protection, 2024.

32 IBM, What is shadow AI?.

33 Reid Blackman, Organizations Aren't Ready for the Risks of Agentic AI, June 13, 2025.

34 IBM, Getting your organization ready to scale responsible agentic AI, April 11, 2025.

35 Karanbir Singh and William Ngu. 2025. Bias-Aware Agent: Enhancing Fairness in AI-Driven Knowledge Retrieval. In Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia. ACM, New York, NY, USA, 8 pages.

36 National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.

37 Béchard, P., Ayala, O. M., Reducing hallucination in structured outputs via Retrieval-Augmented Generation, 2024.

38 Anthropic, Agentic Misalignment: How LLMs could be insider threats, June 20, 2025.

39 Tomorrow Bio, Preventing Bias in AI Models with Constitutional AI, September 21, 2023.

40 Küsters, A., Wörsdörfer, M., Exploring Laws of Robotics: A Synthesis of Constitutional AI and Constitutional Economics, Digit. Soc. 4, 46 (2025).

41 Reworked, Why Agentic AI Projects Fail, July 14, 2025.

42 Bryan Zhang and Kieran Garvey, From Automation to Autonomy: the Agentic AI Era of Financial Services, University of Cambridge Judge Business School, 2025.

43 Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229.

44 Bai, Kadavath, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. Anthropic.

45 IBM, Lack of AI agent transparency risk for AI, July 25, 2025.

46 Microsoft, Audit logs for Copilot and AI applications, August 20, 2025.

47 CIPL thanks Josh Devon for the insights he shared on AI agents and security. Please see his blog for more information.

48 National Institutes of Standards and Technology (NIST), Technical Blog: Strengthening AI Agent Hijacking Evaluations, January 17, 2025.

49 IBM, What is data poisoning, December 10, 2024.

50 Cisco, What is spoofing.

51 Microsoft, What is a DDoS attack?.

52 For example, see HCLTech, Securing AI Agents by Design, accessed August 25, 2025.

53 Salesforce, AI Agents Will Enhance — Not Impair — Privacy. Here's How., February 20, 2025.

54 Sam Sabin, Exclusive: Anthropic warns fully AI employees are a year away, April 22, 2025.

55 Sam Sabin, Microsoft injects AI agents into security tools, March 24, 2025.

56 HCLTech, Securing AI Agents by Design.

57 IBM, Misaligned actions risk for AI, July 25, 2025.

58 Anthropic, Agentic Misalignment: How LLMs could be insider threats, June 20, 2025.

59 For example, see Sean Norick Long, The Mirror Test for AI Agents: A Path to Regulate Autonomous Algorithmic Collusion, September 9, 2025.

60 Colorado Privacy Act Rules, Part 9.

61 HCLTech, Shaping Agentic AI with Responsible AI and Governance, July 2025.

62 CIPL, Q&A on Organizational Accountability in Data Protection, July 3, 2019.

63 CIPL, Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework, February 2024.

64 McKinsey & Company, McKinsey Technology Trends Outlook 2025, July 22, 2025.

65 CIPL, Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework, February 2024

66 Singapore PDPC's Model AI Governance Framework (2nd Ed) features a similar matrix on probability and severity of harm.

67 Sam Sabin, New cybersecurity risk: AI agents going rogue, May 6, 2025.

68 Workday, Workday Agent System of Record.

69 Anthropic, Agentic Misalignment: How LLMs could be insider threats, June 20, 2025.

70 Milvus, AI Quick Reference - How do AI agents handle conflicting goals?.

71 Anthropic, Introducing the Model Context Protocol, November 25, 2024.

72 Google, Announcing the Agent2Agent Protocol (A2A), April 9, 2025.

73 NANDA: The Internet of AI Agents

74 Benjamin Faveri, Craig Shank, Richard Whitt, Philip Dawson, The Need for and Pathways to AI Regulatory and Technical Interoperability, April 16, 2025.

75 CIPL, Regulatory Sandboxes in Data Protection: Constructive Engagement and Innovative Regulation in Practice, March 8, 2019.

76 For example, the ICO has established a Regulatory Sandbox to support organizations creating products and services that use personal data in innovative ways, and the Singapore IMDA has created a sandbox specifically to facilitate experimentation with PETs.

77 Delaware News, Delaware Launches Bold AI Sandbox Initiative, Cementing Its Role as a National Leader in Responsible Tech Innovation, July 23, 2025.

78 CIPL, Learning from Practice: Shaping Effective Sandboxes Under the EU AI Act, September 15, 2025.

79 OECD, G7 reporting framework – Hiroshima AI Process (HAIP) international code of conduct for organizations developing advanced AI systems.