KPMG

AI governance for the agentic AI era

July 2025

kpmg.com

Introduction

We have entered a bold new era in artificial intelligence—one defined by agentic AI—autonomous systems that can perceive, reason, plan, and act with minimal human intervention. These intelligent agents represent a breakthrough in AI-driven decision-making, enabling enterprises to automate complex workflows and adapt in real time.

But this is more than a technological shift—it's a strategic inflection point. As these systems assume more complex, high-value roles across the enterprise, they are not just enhancing operations—they are redefining what's possible. From accelerating decision cycles to enabling adaptive, real-time responses, agentic AI is opening new frontiers for innovation, growth, and competitive advantage.

Yet, as agents become deeply embedded in everyday workflows and take on higher-impact responsibilities, they introduce a new spectrum of risks. The very capabilities that make agents so powerful—their ability to learn, reason, and act independently—also brings ethical, operational, and reputational challenges that organizations must proactively address.

To harness the full potential of agentic AI, organizations must embed trust at the core. This means building systems that are not only intelligent but also accountable, transparent, and aligned with human values. Our Trusted AI framework provides the foundation for this, enabling companies to implement structured risk and compliance programs that mitigate threats while amplifying business impact.

By embracing both the promise and the responsibility of agentic AI, forward-thinking organizations can lead with confidence in an AI-driven future—innovating boldly, operating responsibly, and shaping the next frontier of digital transformation.

Understanding agentic AI through the KPMG TACO Framework™

Agentic AI represents a significant evolution beyond traditional AI. While conventional AI excels at classification, prediction, and pattern recognition, it remains static, requiring periodic retraining and human oversight.

Agentic AI differs by being dynamic and adaptive, capable of making real-time decisions, continuously learning from its environment, and autonomously adjusting to changing conditions. This adaptability makes it particularly valuable for high-stakes business processes where real-time adjustments and advanced decision-making are critical. Companies integrating agentic AI into customer support, software development, and business operations can improve efficiency, enhance workflows, and enhance decision-making.

Our latest AI pulse survey shows agentic AI adoption accelerating far faster than prior AI technologies. Nearly every major enterprise software provider is racing to embed agentic capabilities—a sharp shift from where the market stood even a year ago. As businesses increasingly rely on these systems, strong governance and risk management frameworks will be critical to ensuring agentic AI is deployed responsibly, ethically, and at scale.[1]

[1] KPMG LLP, AI Q2 2025 Pulse Survey (June 2025)

As AI agentic systems proliferate and scale, a structured framework is essential to understand and categorize them based on their capabilities. We can look and assess them by examining the complexity of the goals they fulfil, the depth of planning required, and the level of coordination and orchestration involved.

To make sense of these variations, we created the KPMG TACO Framework – which classifies agents into four key types: Taskers, Automators, Collaborators, and Orchestrators. Each of the types leverages the same foundational tools and capabilities – goal interpretation, reasoning engines (using advanced models including LLMs), memory, tools, and orchestration – but differ in goal planning, execution and complexity.

The four main types of AI agents

Taskers focus on accomplishing singular goals by breaking them into structured, repeatable tasks, making them easy to deploy, scale, and monitor. For example:

- Compliance Agent (Know Your Customer) – Screens customer onboarding documents, detecting compliance risks in regulatory watchlists.

Automators handle goals that require integration across multiple enterprise-wide, automating end-to-end business processes, and cross-functional workflows. Think procure-to-pay, order-to-cash, and compliance auditing. Examples include:

- Healthcare Claims Processing Agent – Manages patient insurance benefits validation, claims adjudication, and payout processing, reducing manual workload and errors.

Collaborators act as AI teammates, working contextually and closely with human operators, enhancing productivity by learning from interactions and refining recommendations. Examples include:

- Marketing Assistant: Brainstorming campaign ideas, refining copy, adapting to feedback

Orchestrators involve multi-agent ecosystems where agents collaborate not only with humans but also coordinate with other AI agents to achieve complex tasks at scale while dynamically adapting to real-time changes. Examples include:

- Supply Chain: Scheduling shipments, managing inventory, and rerouting around disruptions

Mitigating agentic AI risks with the KPMG Trusted AI framework

As we push the boundaries of what AI can do, we must confront not only technical challenges, but also deep ethical and societal questions about control, accountability, and trust.

The KPMG Trusted AI™ framework offers an actionable approach to managing these risks in a responsible and ethical manner. It equips organizations with the tools to embed ethical principles and governance into every stage of the AI lifecycle—from design and deployment to monitoring and evolution. Below are key highlights of how this framework can be applied to help ensure trusted agentic AI deployments.

Reliability

One of the key emerging risks associated with agentic AI systems is misalignment—when the system's actions diverge from the human's original intent. This misalignment can significantly undermine reliability, especially when objectives are not clearly defined or oversight is lacking.

According to the KPMG Trusted AI framework, reliability refers to the extent to which AI systems perform consistently with their intended purpose, scope, and required level of precision. Reliability requires not only robust design and testing but also continuous monitoring to align outcomes with human expectations and values.

Accountability

Accountability helps ensure that organizations clearly define responsibility for AI-driven decisions and that there is an audit trail of agent activity. As AI agents become more autonomous, assigning accountability and enabling traceability becomes increasingly complex but important. This ambiguity can heighten legal and ethical risks, particularly when AI systems produce unintended or harmful outcomes.

A contributing factor is automation bias—the human tendency to place excessive trust in AI outputs, especially in high-stakes situations. As oversight diminishes, organizations may rely too heavily on AI agents to perform complex tasks without adequate validation or human intervention.

The KPMG Trusted AI framework addresses these challenges by advocating for clearly defined roles and responsibilities across the AI lifecycle and encouraging human-in-the-loop. This includes developers, users, and other stakeholders, ensuring that accountability is embedded from design through deployment and ongoing use.

Transparency & Explainability

Maintaining transparency and explainability in agentic AI systems is increasingly challenging due to their dynamic, adaptive behavior and complex, multi-step reasoning processes. These systems often evolve over time and operate across interconnected workflows—sharing data, delegating tasks, and coordinating strategies. In such environments, even minor errors can quickly propagate, triggering cascading failures throughout the AI supply chain.

According to the KPMG Trusted AI framework, transparency and explainability require agents to clearly communicate why actions are taken, including confidence thresholds, guardrails, and human oversight. This includes interfaces that support interpretability, traceable inter-agent handoffs, and upfront declarations of agent capabilities and limitations. These elements should be reflected in recurring AI system cards to ensure trust and accountability.

Security & Safety

Agentic AI systems introduce heightened security and safety risks by expanding the attack surface and enabling more sophisticated forms of misuse. Their autonomy, access to tools, and ability to operate without direct human oversight make them attractive targets for bad actors seeking to manipulate agent behavior, poison agent goals, or exploit system vulnerabilities. These risks are especially pronounced when agents interact with external systems or operate across organizational boundaries.

According to the KPMG Trusted AI framework, security involves implementing robust and resilient practices to protect AI systems from unauthorized access, manipulation, or disruption, while safety focuses on preventing emotional and/or physical harm to people, businesses, and property. To address these concerns, the framework recommends designing AI systems with layered security controls, built-in fail-safe switches, and rollback protocols—reinforced by continuous AI red-teaming efforts to proactively test vulnerabilities and uphold a "trust but verify" approach.

Data Privacy

Agentic AI systems often process significantly larger volumes of data in various forms to support their reasoning, memory, and decision-making capabilities. This increased data flow raises the risk of privacy breaches, especially when data is shared across systems or used in unintended ways. AI agents are also more susceptible to indirect prompt injection attacks through subtle instructions, allowing attackers to gradually extract sensitive data without breaching systems directly. Their ability to operate across platforms can unintentionally link data, while inference attacks exploit patterns in agent responses to uncover confidential information.

The KPMG Trusted AI framework addresses these concerns by promoting strong data governance practices, including the use of anonymized and ethically sourced data, strict access controls, and compliance with data protection regulations. These safeguards help ensure that even as data volumes grow, individual privacy remains protected.

Fairness

AI agents also raise fairness concerns, a key focus of KPMG's Trusted AI framework, which stresses the importance of minimizing bias against individuals, communities or groups. These systems often adopt personas—structured behavioral profiles designed to guide interactions—which can unintentionally encode and perpetuate biases rooted in their training data, model architecture, or in the way agent instructions and feedback are written. When such biases are embedded in the agent's persona, they don't just influence responses, they shape decisions. Unlike passive systems, agentic AI can act based on this skewed logic, potentially reinforcing unfair outcomes across a wide range of interactions. To address these risks, the framework encourages organizations to embed fairness metrics and thresholds into agent design, data sources and governance, supported by continuous evaluation and feedback mechanisms.

The first 10: Building a trusted foundation for agentic AI

Aligned with the KPMG Trusted AI framework, these top 10 control considerations serve as a powerful foundation for deploying agentic AI responsibly and effectively. This is not a one-time checklist—it's a continuous, evolving journey. Each step forward strengthens trust, transparency, and resilience.

1 Assess agent risk

- Complete an impact assessment and determine the risk level of the agent
- Risk assessments can help determine levels of human review and other mitigation tactics

2 Determine human oversight requirements

- Define the stages within the decision-making and action process where human review and/or approval is required prior to execution, based on risk assessment
- Establish oversight mechanisms including human-in-the-loop, human-on-the-loop, or machine-in-the-loop configurations

3 Establish default scope boundaries

- Provide agent with baseline objectives, guiding principles, values, and responsible AI frameworks that allow actions without guidance
- Define clear scope constraints on what an agent can and cannot do, including limits on data access, tool usage, and decision-making authority
- Require agents to flag ambiguous situations and escalate to human review when needed

4 Reveal agent's chain-of-thought thinking

- Design agents to reveal intermediate steps, assumptions, or data and logic used to arrive at conclusions—especially for complex tasks or high-impact decisions
- The level of detail should be appropriate to the audience (e.g., technical teams vs. end users) and the task (e.g., critical decisions vs. routine automation)

5 Assign unique identifiers for attributability

- Assign unique identifiers to agents and humans within AI systems and workflows to help trace all actions and decisions

6 Design immutable logging and monitoring

- Implement an immutable audit trail of agent interactions and decisions
- Establish ongoing monitoring, thresholds on key metrics, and alert mechanisms so anomalies can be detected

7 Design multi-agent systems to help prevent cascading failures

- Identify integration points between AI systems to detect potential chain reaction triggers
- Implement mechanisms that automatically halt AI operations across systems, when unexpected behavior is detected

8 Build fail-safe and fallback protocols

- Design a fail-safe switch to turn an agent off depending on pre-defined thresholds
- Define fallback protocols when agents are shut down
- Design solutions so agents can not interfere with fail-safe mechanisms

9 Deploy AI red-teaming

- Enforce technical controls such as meta prompting to validate and block manipulative prompts
- Design guardrails to sanitize all inputs and outputs to prevent unauthorized actions or data leaks
- Deploy AI red and purple teaming to continuously test agents for trustworthiness in alignment with an organization's responsible and ethical principles

10 Ongoing evaluations & feedback

- Define pre-launch and ongoing evaluation benchmarks. Set measurable thresholds for key metrics such as repeatability, accuracy, safety, fairness, and data privacy
- Help ensure that the evaluation is completed at the component (sub-task/decision) or tool/orchestration level (overall action/decision) testing
- Establish recurring schedules to periodically provide feedback, reinforcement learning, and retraining of agents to prevent drift

Conclusion

The agentic wave is here, marking a pivotal moment in business transformation. This leap forward in AI capability isn't just changing the game—it's creating an entirely new playing field where autonomous systems drive unprecedented value and competitive advantage.

Success in this new era demands more than just technological prowess. It requires a deliberate approach that places trust at the core of AI deployment. Organizations must build their agentic AI foundations on robust governance frameworks that enable innovation while helping to ensure responsible implementation. The KPMG Trusted AI framework provides this critical foundation, enabling businesses to accelerate transformation while maintaining control and building stakeholder trust.

Are you prepared to lead the charge with agentic AI? The future is here—dynamic, powerful, and full of promise. Seize it.

Reach out to the KPMG experts who will help you mitigate risks within your AI systems.

Bryan McGowan
Principal,
KPMG Global and
US Trusted AI Lead
bmcgowan@kpmg.com

Aisha Tahirkheli
Managing Director,
US Trusted AI
atahirkheli@kpmg.com

Kartik Gupta
Senior Manager,
Trusted AI,
KPMG in Canada
kartikgupta3@kpmg.ca

Kareem Sadek
Partner,
Advisory Tech Risk and
Trusted AI Lead,
KPMG in Canada
ksadek@kpmg.ca

Nana Amonoo–Neizer
Director,
KPMG Global and
US Trusted AI
namonooneizer@kpmg.com

Brad Wrobleski
Manager,
AI Advisory,
KPMG in Canada
bwrobleski@kpmg.ca

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

Learn about us: **kpmg.com**