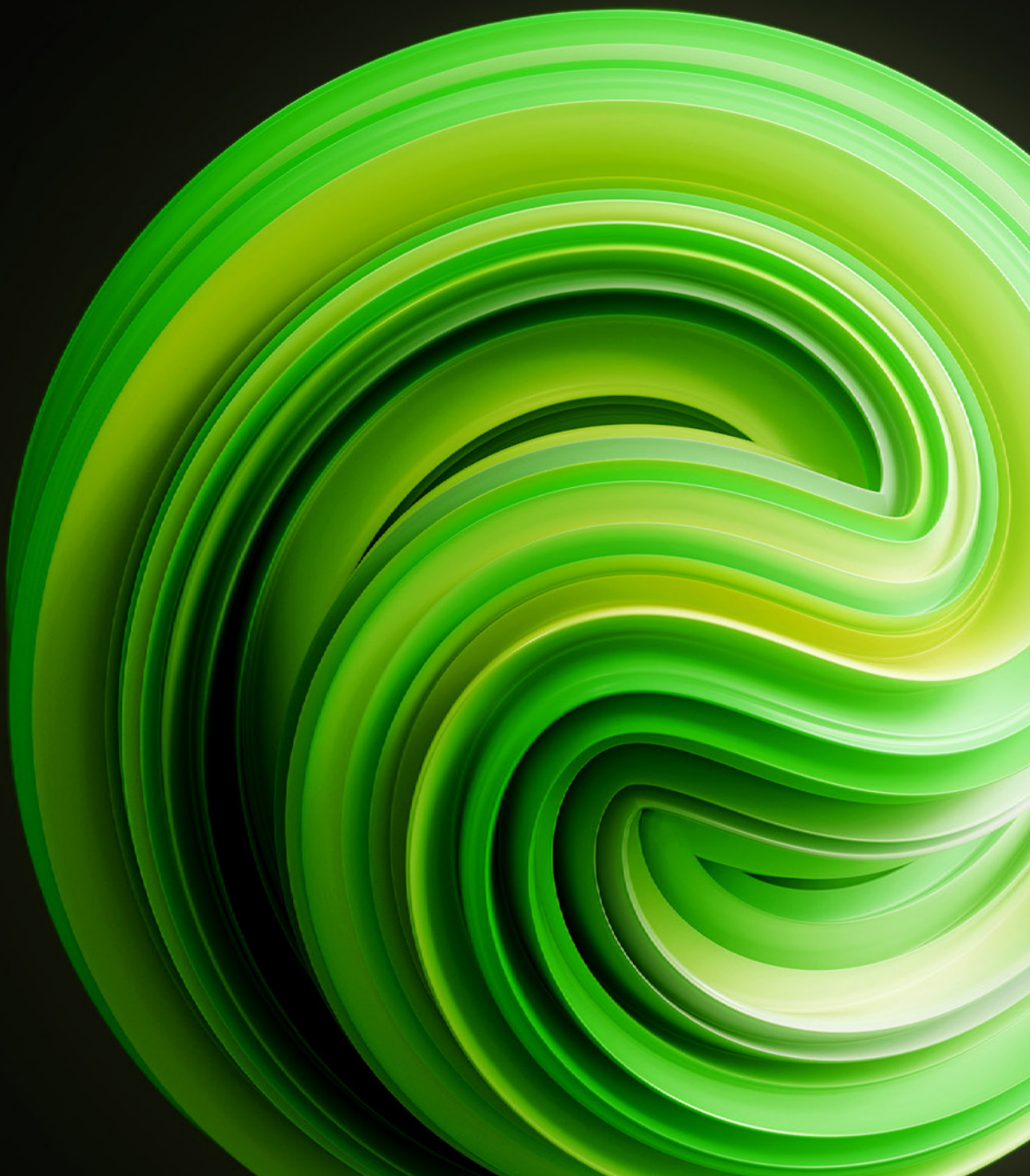
Deloitte.

From hype to control

Validating Agentic AI



Agentic AI is powerful and autonomous, but trust isn't automatic.

This whitepaper shows you how to validate, monitor, and stay in control.

Contents

Key takeaways

1

Agentic AI: transformative potential meets complex challenges

Agentic AI offers significant business transformation through autonomous operations and complex workflow management. However, this comes with inherent challenges.

2

The imperative for a strategic, holistic validation framework

Successful adoption of Agentic AI demands a proactive and continuous validation approach, moving beyond reactive, pre-deployment checks.

3

Continuous monitoring and human oversight for trust and performance

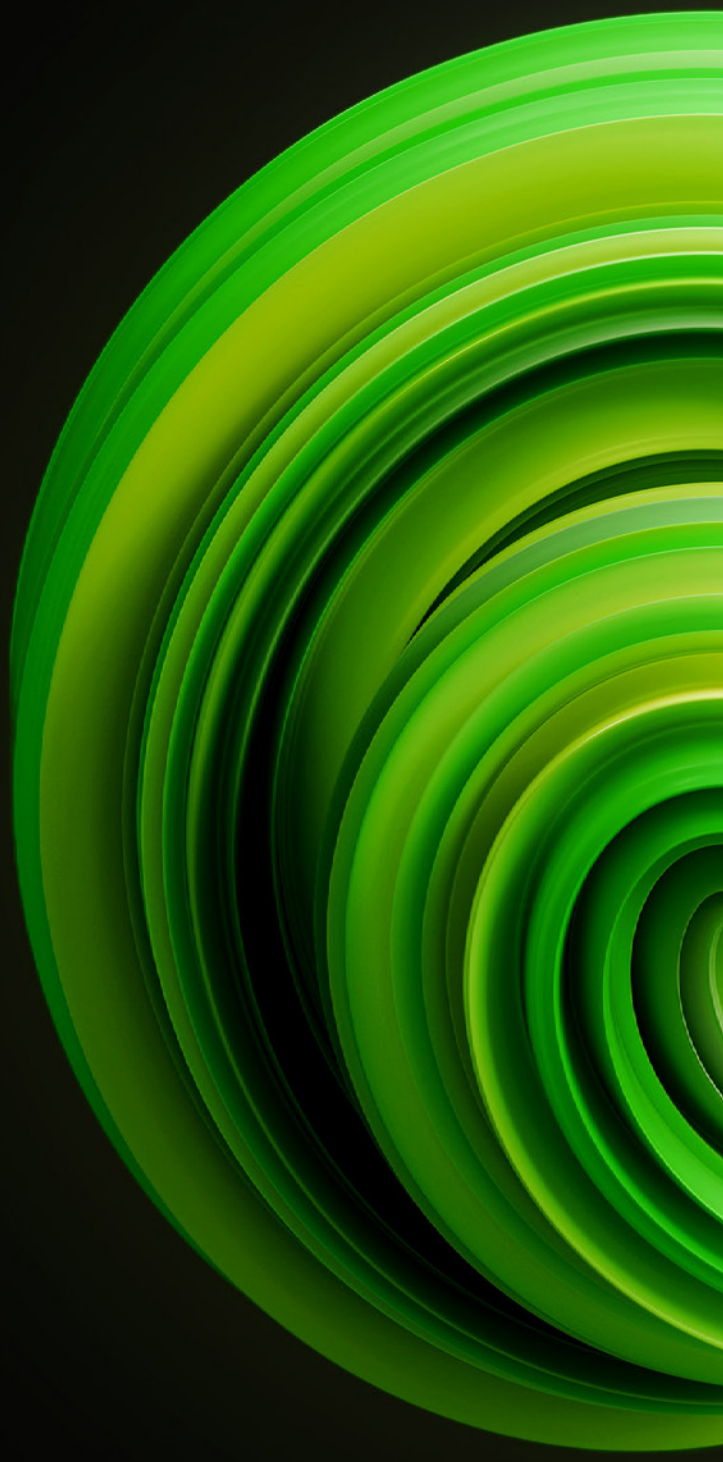
Given the dynamic and autonomous nature of Agentic AI, continuous monitoring, real-time observability, and human-in-the-loop validation are essential.

Introduction

The business landscape is undergoing a transformation driven by the rise of Agentic AI – systems capable of independent operation and complex workflow management at scale.

Unlike traditional AI, Agentic AI has different levels of autonomy enabling it to execute tasks with minimal or no human oversight, which is fundamentally altering how organisations automate processes and make decisions.

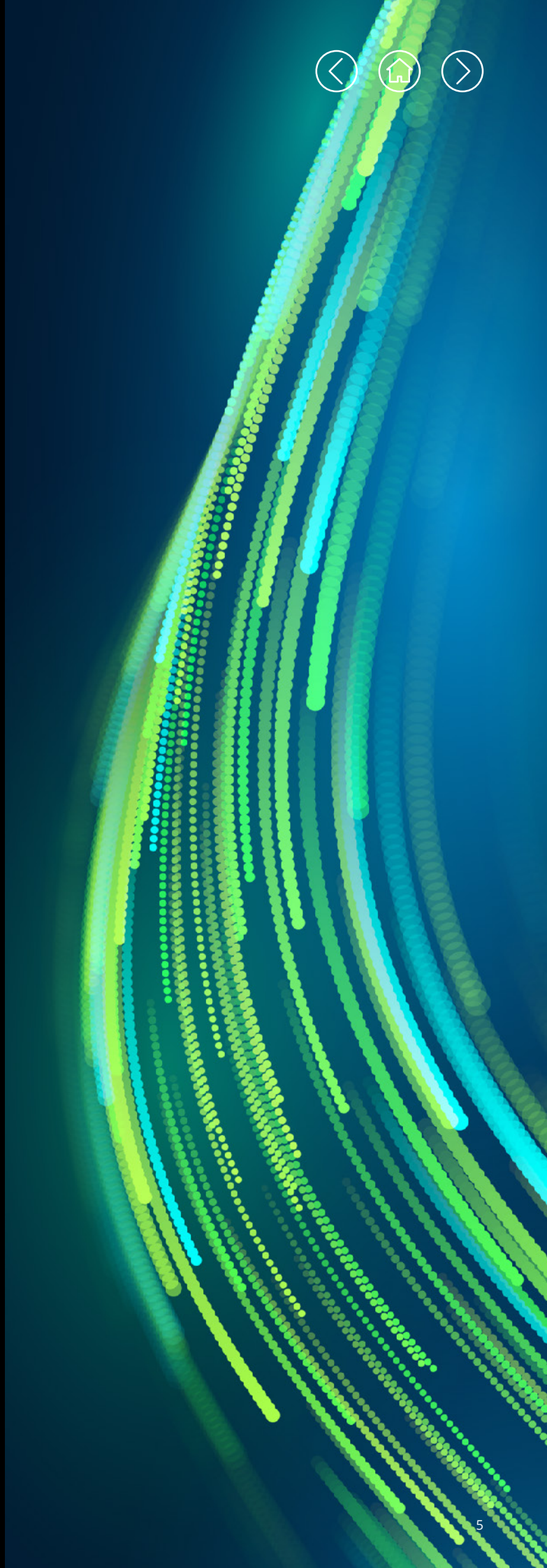
However, with this autonomy comes a critical challenge: how can businesses ensure these powerful, autonomous systems are reliable, safe, and consistently aligned with their strategic objectives?



For successful adoption, this whitepaper emphasises the need for a strategic, proactive approach to Agentic AI validation, moving away from reactive, pre-deployment checks toward continuous, integrated practices within both business and technical operations.

Given the intrinsic complexity of agent autonomy and unpredictable behaviour, a comprehensive framework is essential. This must cover governance, regulatory compliance, performance, accuracy, safety, and ethical considerations.

By embedding validation as an ongoing process, businesses can effectively manage autonomous systems and secure a leading market position in the era of Agentic AI.



Understanding Agentic AI in a business context

Early enthusiasm for Large Language Models (LLMs) and their potential to transform business and process automation was somewhat misplaced as they proved difficult to integrate into the real-world without significant human involvement.

Agentic AI is now unlocking a significant wave of exciting new possibilities, enabling more sophisticated transformation of business processes. It is a significant departure from traditional models, such as decision trees or early neural networks.

While traditional AI is trained within predefined parameters for specific tasks, it often lacks adaptability and contextual memory. Agentic AI systems are autonomous entities designed to pursue goals independently. Agentic AI can reason, plan an execution strategy, choose and use appropriate tools (like calling an API or using a search engine), and adapt its actions based on real-time feedback to achieve its objective (Figure 1).

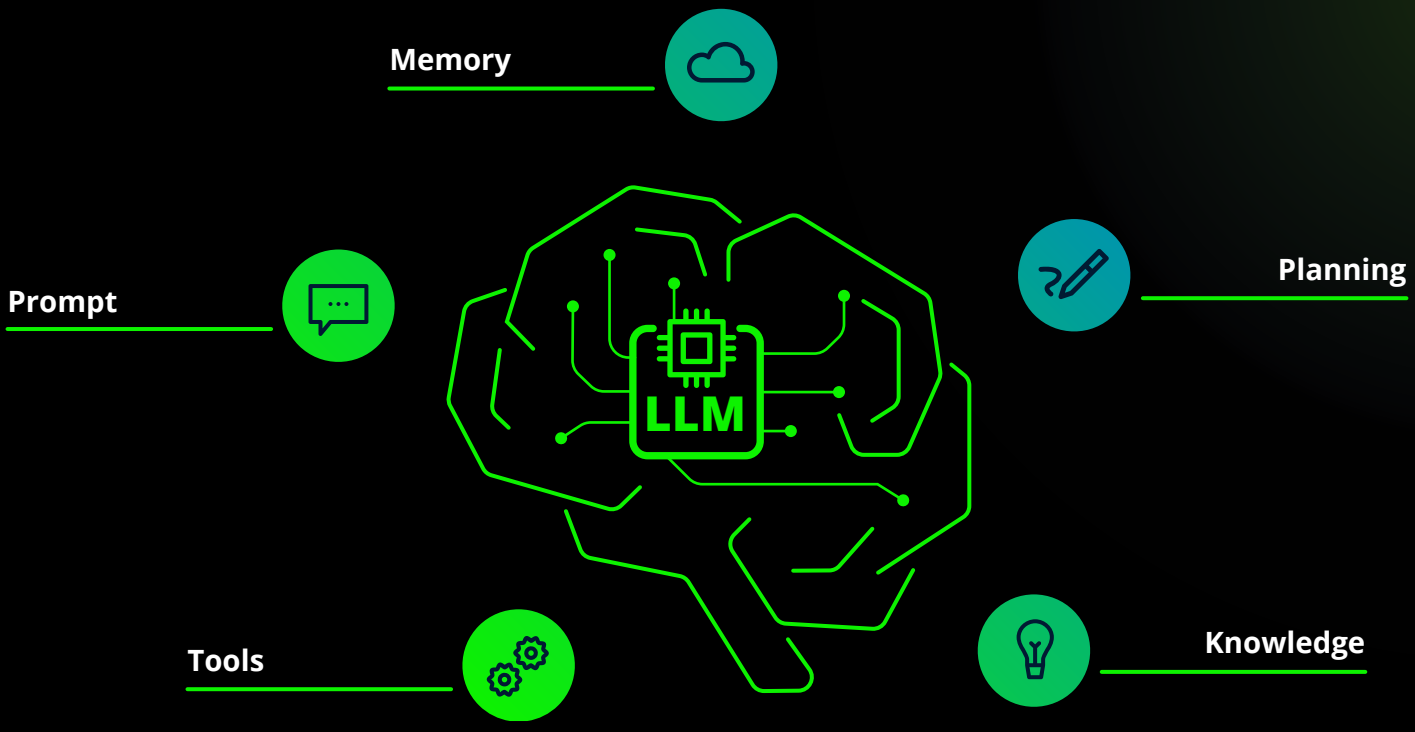


Figure 1: Key components of Agentic AI

While the terms "Agentic AI" and "AI agents" are often used interchangeably, it is critical to understand the subtle difference. An AI agent is a single software component/entity designed to execute a specific task, such as a chatbot following a set script. While autonomous, it is largely task-specific and remains bound by its fixed, encoded logic, unable to self-improve or dynamically adapt[1].

Agentic AI, however, is the overarching, goal-driven system. It acts as an orchestrator, deploying and coordinating multiple AI agents as building blocks to achieve a complex, ultimate outcome that typically requires minimal human oversight[2]. Examples of AI Agent capabilities that can be part of Agentic AI systems are shown in Table 1.

Table 1: Key capabilities of autonomous AI Agents

Natural Language Reasoning	LLM agents understand context, interpret ambiguous instructions, and reason through complex problems using natural language as their primary interface.
Tool Use and Function Calling	Modern AI agents, including LLM, can invoke external tools, APIs, and services (e.g. search the web, execute code, query databases, etc.)
Multi-Step Planning	Agent can decompose complex objectives into subtasks, creating and executing plans that might involve multiple steps.
Memory and Context Management	Advanced agents maintain conversation history, store relevant information for future use, and manage their own context windows to handle long-running tasks that exceed single prompt limitations.
Code Generation and Execution	AI agents can write, debug, and execute code, enabling them to perform computational tasks, data analysis, and even self-improvement through code generation.

Agentic workflows are iterative, using a feedback loop of thinking, researching, and revision to manage complex tasks more efficiently. This enables them to execute real-world actions, such as sending emails or querying databases via external tools, and continuously refine strategies.

The enterprise-wide adoption of Agentic AI is expanding, with diverse applications across functions. "Individual augmentation" is currently the most common use case, enhancing employee capabilities and productivity through tools like smart assistants and content generators for tasks such as document processing, and information retrieval[3].

Another growing trend is "workflow automation", which automates and optimises business processes by orchestrating complex tasks across multiple steps – and even systems – and monitoring performance. Depending on the Agentic AI use case, different types of AI agent are typically implemented, with their different patterns shown in Table 2.

Example of design patterns for AI agents and applications

Reflection

The agent refines its subsequent actions based on self-evaluation of outputs.

Examples:

- Code generation and review.
- Writing and refining drafts.
- Problem solving and planning, evaluating feasibility.
- Complex problem solving.

Tool Use

Agents can invoke external functions, retrieve information, or perform actions.

Examples:

- Retrieving information from websites via APIs
- Personal assistant
- Searching vector documents

Multi-agent

Several distinct agents, each with a specific role, collaborate towards a shared objective.

Examples:

- Brainstorming with different "personas"
- Complex software creation
- Running virtual simulations

Planning

An orchestrator solves complex problems by breaking down plans into subtasks.

Examples:

- Multi-modal tasks involving images, text and data
- Software development plans
- Research and report generation

Table 2: Different patterns of AI Agent within Agentic AI[4]

Agentic AI is proliferating across numerous sectors, streamlining supply chain processes[5], enabling hyper-personalisation in customer engagement[6], and facilitating predictive risk management[7].

The true value of an Agentic AI lies in its ability to solve real business problems and drive efficiency, innovation, and competitive advantage. As businesses focus on high-value AI applications, rigorous validation and governance of these autonomous systems is becoming paramount to making sure business scale safely and achieve their desired ROI.

Core Challenges with Agentic AI

1. Beyond the Algorithm: Navigating the Hurdles of Real-World Deployment

Harnessing the benefit of Agentic AI comes with some important challenges including navigating the unpredictable actions of autonomous systems, understanding how AI arrives at its decisions, mitigating the risks of over-reliance on automation, and addressing crucial ethical considerations[8] (Figure 2).

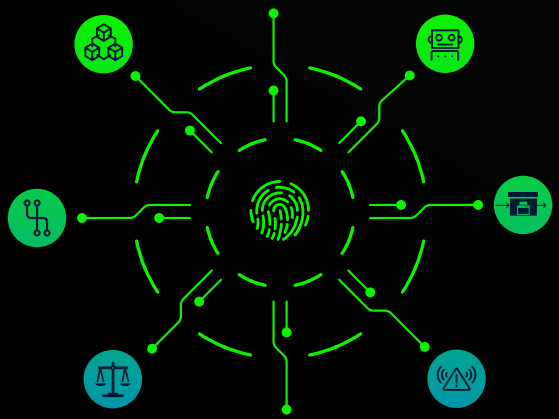
A robust and holistic approach to managing the inherent complexities of Agentic AI, that is underpinned by key principles of Trustworthy AI, is therefore essential. This begins with establishing strong enterprise-level governance frameworks to align AI systems with business objectives and regulatory requirements. It also necessitates ensuring comprehensive legal and regulatory compliance, addressing AI-specific laws, data protection, and sector-specific mandates. Rigorous Technical validation and continuous monitoring of Agentic AI systems are essential to ensure they are robust and reliable, responsible, fair and impartial, accountable, transparent and explainable.

Finally, securing the underlying platform and infrastructure is paramount to ensure they are private, safe and secure which involves a thorough evaluation of data protection, system vulnerabilities, and third-party considerations.

Agentic AI Integration
Complex integration into existing IT environments, demanding robust solutions for data governance, security, and autonomous decision-making

The Risk of Over-Autonomy
Relying too heavily on AI can sideline essential human judgment and erode valuable skills, highlighting the need for balanced oversight to ensure safety and effectiveness.

Unpredictable Behavior
AI agents can deviate from their original programming as they learn (behavioral drift), causing unpredictable actions and consequences that may violate compliance or create new risks.



The "Black Box" Challenge
The difficulty in understanding how AI models arrive at their decisions poses a major hurdle for accountability, trust, and proving compliance to regulators

Fairness, Bias, and Regulatory Compliance
Agentic AI poses ethical challenges around bias and privacy; AI system must operate fairly, without bias, and in constant alignment with a complex and evolving regulatory & legal landscape.

Risk Management & Continuous Monitoring
Effective AI deployment requires constant supervision to manage risks, monitor for performance degradation ("model drift"), and protect against security threats.

Figure 2: Key challenges of implementing and validating Agentic AI systems

2. Technical and Operational Considerations

Deploying Agentic AI can be complex, especially with legacy system integration[9], which can impact information processing and costs. Therefore, careful evaluation of scalability and overall cost is crucial to ensure a positive return on investment. Furthermore, the "black box" nature of some Agentic AI systems complicates troubleshooting, necessitating advanced monitoring and explainability.

2.1 Agentic AI Lifecycle Management:

Proactive lifecycle management is crucial as Agentic AI must be able to handle large datasets and high user volumes without performance degradation. Performance can deteriorate over time due to "Agentic/behavioural drift", due to the dynamic, unpredictable nature of some AI agents[10]. Therefore, ongoing maintenance and updates are required to prevent the Agentic AI system from showing unpredictable behaviour or failing to handle exceptions.

2.2 Explainability and the "Black Box" Problem

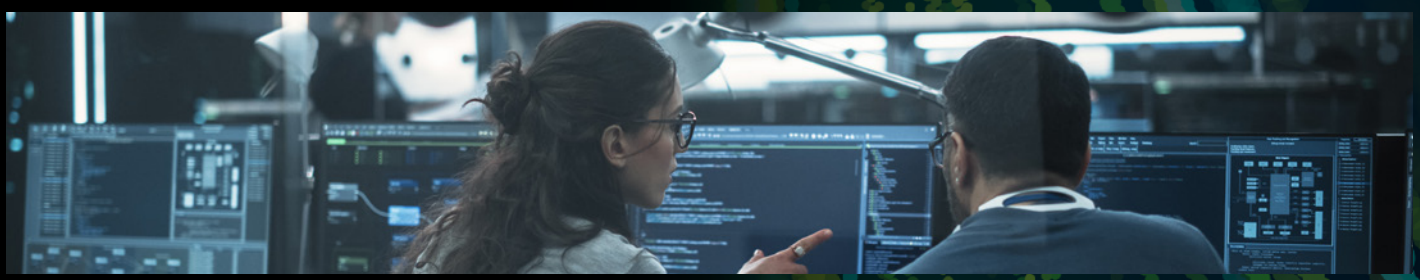
The inherent complexity and opacity of Agentic AI systems, known as the "black box" problem, is perceived as a potential challenge to trust and adoption, particularly in regulated sectors. Their autonomous and iterative decision-making processes deepen this issue considerably[11]. Articulating why an AI agent acted or concluded something is fundamental for troubleshooting, continuous refinement, and compliance[12]. Explainability is not just about building trust but is a foundational requirement for decision transparency and a non-negotiable legal and compliance necessity in sectors like finance and healthcare[13].

3. Regulation and Governance Challenges:

Deploying Agentic AI presents considerable regulatory compliance and governance challenges, fundamentally driven by the contradiction between legislative demands for auditable predictability and the Agentic AI systems' inherent capacity for complex, autonomous and emergent behaviours. Organisations must navigate an evolving, fragmented global landscape[14 15], which necessitates adherence to both prescriptive, risk-based mandates, such as those found in the EU AI Act[16], and more principles-based approaches like that in the UK.

Demonstrating compliance requires moving beyond static testing to implement stringent, verifiable technical assurance methodologies that continuously validate trustworthy AI characteristics, including accuracy, robustness, safety, explainability and transparency, and bias mitigation, across dynamic, multi-agent interactions[17 18].

Consequently, effective oversight and risk management frameworks must adopt flexible governance models, establish clear accountability, and maintain robust decision-making trails and flexibility to accommodate continuous learning and evolution. It must also ensure Agentic AI technical assurance methodologies meet or exceed the highest global requirements to secure compliance with existing – and any relevant future – regulations.



4. Security and Privacy Challenges:

The autonomous nature of Agentic AI introduces security and privacy challenges too. Their potential deep access to sensitive data and other systems' functionalities makes them prime targets for cyberattacks[19], leading to the potential for compromise of confidential information. Potential vulnerabilities exist across their communication networks too[20]. AI Agents are susceptible to prompt injection attacks, where malicious inputs manipulate behaviour, potentially leading to data exfiltration or misuse of tools[21]. Furthermore, agents' autonomy in data collection and retention can increase the risk surface and complicate compliance with privacy rights like the "right to be forgotten"[22]. This is particularly challenging because Agentic AI systems could access and utilise unclassified sensitive information that has been wrongly filed (such as sensitive data stored incorrectly in platforms like SharePoint) without human knowledge, making it impossible to know when a data breach has occurred.

5. Ethical Challenges

The autonomy of Agentic AI brings significant ethical challenges. At the forefront is algorithmic bias, where the system can inherit and amplify biases from the data, leading to discriminatory outcomes and tangible harm[23]. In high-stakes environments, misjudgements pose significant safety risks, raising critical questions of accountability. Other crucial ethical considerations include the explainability of the decisions and ensuring that the Agentic AI's goals truly align with human values and intended outcomes.

Defining and measuring fairness is complex, varying across cultural, legal, and ethical contexts[24]. Organisations must develop context-aware frameworks, extending beyond technical checks to include legal and ethical considerations from the outset, and continuously test for and mitigate bias.

Core validation domains

Validating Agentic AI is a developing area that demands a comprehensive, bespoke approach to evaluate performance across numerous dimensions. The unique characteristics and architecture of Agentic AI make validation highly complex and often application-specific. It is crucial to assess both the system as a whole and its individual components[25]:

Agentic AI Architecture Assessment

Given the complexity of Agentic AI systems, a well-designed architecture is crucial for optimal performance. This demands a thorough examination to ensure a logical and efficient system flow. Key areas for assessment include:

- **Decision-making architecture:** An assessment of where and how the Agents/LLM(s) are positioned within the overall process flow. This is especially important for multi-agent systems where each agent/LLM may have a specialised role in the application's overall process flow. Task-splitting and positioning in this case will be crucial to performance.
- **Tool availability at each step:** This ensures sufficient and appropriate tools are available for each step, and that each LLM/agent receives the necessary inputs to use these tools correctly.
- **External connections:** For systems with external inputs or connections (e.g., databases, web links), this assessment verifies that these connections are correctly configured.
- **Load/scale testing & error recovery and resilience:** Testing the system performance under various load conditions. This includes testing with concurrent requests, long-running tasks, and resource constraints. In addition, testing how the system handle various failure scenarios and ensure the system attempt reasonable recovery strategies.

Component-by-Component Validation

Agentic AI systems typically comprise multiple components, each handling specific tasks. The nature of these components varies depending on the application. Multi-agent systems may include specialist AI agents, while single-agent systems might incorporate LLM-related sub-applications (e.g. retrieval augmented generation (RAG) or code generation). Evaluating these components individually ensures efficient and intended system operation. Evaluation methods vary depending on component function, but common patterns include assessing reflection, tool use, and planning capabilities[26 27]. A simple single-agent system might use one set of these components, whereas a multi-agent system may employ several.

When validating Agentic AI systems, whether validating its system architecture or its individual components, the following dimensions are essential considerations to ensure a comprehensive approach:

1. Performance and Efficiency

This dimension measures the operational effectiveness of the AI agent in a production environment. Key metrics may include:

- **Success Rate/Task Completion:** The proportion of tasks or goals that the AI agent completes correctly without human intervention. When measuring success rate, the system's output to a given user request can be evaluated for correctness and quality, using standard LLM metrics like response relevance and hallucination rates. This user-centric assessment can be termed "user-success"; a simple pass/fail based on the user's perspective[28]. However, the AI agent's decision-making process itself must also be evaluated against predefined criteria, which we can call "system-success".
- **Error Rate:** The percentage of incorrect outputs or failed operations.
- **Latency:** The time taken for an AI agent to process a request and return a result.
- **Cost:** Measures resource usage, such as token consumption or compute time, which can be a primary driver of operational costs.

While evaluating performance, it important to perform **scenario testing** where Agentic AI's performance is evaluated across various use cases and scenarios to ensure accurate responses and request fulfilment. Testing should include both common situations (assessing typical performance) and challenging scenarios (acting as a stress test for the system)[29].

Furthermore, as **tool use** is a critical aspect of Agentic AI, it is important to monitor and validate the tool use. Monitoring tool use involves various methods and metrics, often tailored to the specific application. For simple AI agents not reliant on external factors, simple metrics that check of whether the AI agent's tool calls adhere to a user-defined ideal sequence might be sufficient.

However, more complex AI agents interacting with users over multiple turns or depending on external conditions (e.g., internet availability) require sophisticated systems to simulate these conditions during evaluation[30].

2. Accuracy and Reliability:

For Agentic AI systems, accuracy is a nuanced concept. It is not a simple pass/fail metric but a combination of multiple components, that is highly dependent on the use case;

When assessing accuracy for example, **factuality** (is the system retrieving factually correct data?) and **relevance** (does the system present relevant insights that directly answer the question?) could be equally important[31]. **Instruction following accuracy** tests whether AI agents accurately interpret and execute instructions across a range of phrasing and complexity. This includes verifying their understanding of context, their ability to manage ambiguity, and their capacity to request clarification when necessary. **Reasoning chain tests** examine the AI agent's reasoning process and whether the logical steps make sense, assumptions are reasonable, and conclusions follow from premises.

The non-deterministic nature of AI agents means their performance can vary based on factors like prompt phrasing or underlying model updates. Therefore, reliability can be evaluated by establishing the consistency of the AI agent's performance over time and across different scenarios. **Determinism testing** for example, test that that agents produce functionally equivalent outputs for similar inputs. **Temporal consistency** tests how agents maintain consistent behaviour over extended interactions; where information are not contradicted later, and decisions remain logically consistent.

3. Decision Trajectory

In response to a request, Agentic AI systems determine and execute a series of steps to fulfil the task. This sequence of the steps that Agentic AI takes is known as **decision trajectory**. These steps might involve querying tools, routing the request to specialist AI agents within the system, or seeking further information. Assessing the decision trajectory is crucial for validation, ensuring the Agentic AI's actions remain aligned with the overall goal[28, 32].

Assessing the decision trajectory often involves comparing the Agentic AI's actual steps to an ideal or expected sequence. If the actions match the ideal trajectory, the test is passed[26]. However, as multiple paths may achieve the same outcome for a given request, alternative validation approaches might be needed to focus on "**milestones**" (required steps) and "**minefields**" (steps that must be avoided) within the decision trajectory to determine test success[30]. The selection of which method to use will depend on the complexity of the Agentic AI application. Some decision trajectories may include repeated steps. While sometimes necessary, excessive repetition suggests inefficiency. Therefore, the **frequency** of repeated actions is often monitored during evaluation.

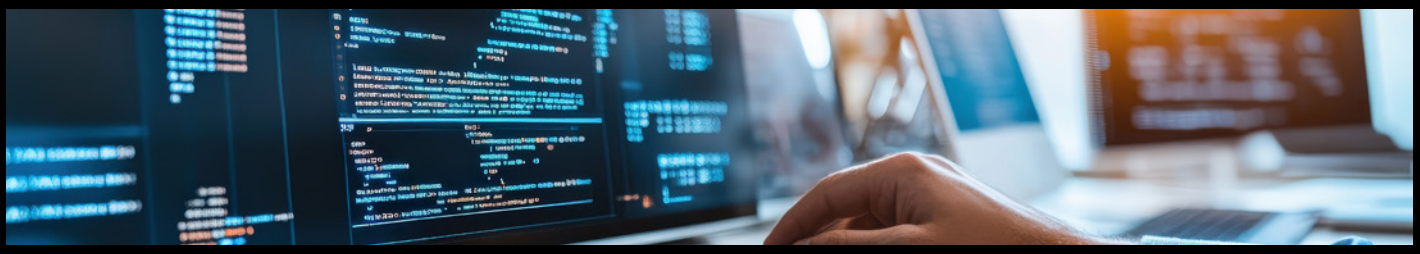
4. Safety and Security:

Validating the safety and security of Agentic AI systems is crucial, especially given their access to sensitive data and systems. This requires rigorous testing to ensure data protection, compliance with data privacy regulations, and resistance to unauthorised access or manipulation[33]. A key element is **AI red teaming**, using simulated adversarial attacks, potentially automated with specialist tools, to uncover vulnerabilities before deployment.

This includes checks for content safety risks, jailbreaks, and other adversarial probing techniques. Further validation involves testing the effectiveness of **safety guardrails** against harmful behaviours (including those concerning sensitive topics and unauthorised operations), assessing resistance to jailbreak attempts (prompt injection, instruction override, social engineering). **Bias and fairness testing** should be included to evaluate the system's outputs for biases related to demographics, cultures, or other protected characteristics, and to validate the alignment with organisational values and policies (ethical decision-making, privacy, and adherence to business rules). Implementing mechanisms to detect and measure **hallucination** is also paramount to ensuring the LLM/AI agent acknowledges uncertainty and avoids fabricating information[34].

5. The Role of Human-in-the-Loop Validation:

Human-in-the-Loop (HiL) is a crucial best practice, combining automation with human oversight to ensure critical decisions align with human judgment and business context. Human intervention at **key stages** validates decisions, handles exceptions, and provides final confirmation, especially in high-stake industries. This isn't solely for error correction; it creates a continuous feedback loop, enabling the AI system to learn and improve. HiL is essential for managing the inherent unpredictability of Agentic AI and fostering user trust[35].



Agentic AI Validation Approaches – Current Landscape

While there is consensus on some key validation features for Agentic AI, a concrete, unified framework for evaluation (i.e., a standardised setup and process) remains elusive.

Different institutions and academics propose varying methodologies. Table 3 below is a summary of the two most common approaches currently proposed:

Table 3: Summary of Agentic AI Validation Approaches

Approach	Methodology	Key features	Strength	Weakness	Organisation
Component-Wise Validation	Breaking down the Agentic AI application into individual steps and evaluating each in isolation.	• Focuses on individual LLM/AI Agent capabilities (Skills) and should be evaluated in isolation. • Monitors decision paths and reasoning, specifically monitoring for path errors and iterations.	Maximises coverage, allows for targeted improvements.	Can be less robust to variations in decision trajectories across different tasks.	Arize AI[27] ORQ[26] IBM[36]
Holistic Assertion-Based Validation framework	Monitors decision paths and reasoning, specifically monitoring for path errors and iterations.	• Defines "Milestones" (required events) and "Minefields" (events to avoid). • Uses scenarios (challenging and common use cases). • Automatic evaluation with an evaluator LLM.	More robust to variations in decision trajectories.	Requires significant upfront design and scenario creation.	Apple30 AWS[28] NEC laboratories[32]

The reviewed approaches reveal a lack of a single, universally accepted validation framework for Agentic AI. However, a hybrid methodology combining component-by-component and assertion-based (system level) frameworks offers a promising solution. Component-by-component validation, analogous to unit testing, would ensure each individual step functions correctly and efficiently.

The assertion-based framework, mirroring integration testing, would then assess the overall decision trajectory, AI agent coordination, and high-level system performance, verifying the effective integration of the individual components.

The Emergence of Real-time and Continuous Monitoring

As Agentic AI transitions from experimental sandboxes to mission-critical business operations, the validation landscape is shifting from static, pre-deployment "snapshots" to real-time observability and continuous monitoring. The market is observing the emergence of continuous monitoring platforms providers, such as Galileo[37], LangSmith[38], Arize Phoenix[39], which are designed to provide visibility into the "black box" of Agentic reasoning. Their primary purpose is to move beyond simple uptime metrics to evaluate decision quality. These platforms typically provide granular traceability and chain-of-thought analysis, mapping every action, tool call, and reasoning steps to help developers pinpoint exactly where a multi-step workflow may have deviated.

Furthermore, they incorporate real-time guardrails, by utilising high-speed 'judge' models or policy managers, to intercept and block hallucinations, prompt injections, or policy violations before they reach the end user. This is complemented by enterprise-grade drift and failure detection, which systematically identifies nuanced issues such as tool misuse or goal drift that often remain undetected during initial benchmarking.

The surge in popularity of these frameworks is driven by the inherent complexity of Agentic systems, and the potential for "cascading errors" where one minor reasoning flaw could lead to a material operational mistake. Businesses are increasingly realising that continuous monitoring is the "insurance policy" required for a successful Agentic AI deployment. Firstly, continuous monitoring facilitates robust risk mitigation by intercepting unauthorised actions, biased outputs, or data leaks in real time. Beyond security, continuous monitoring enables rigorous cost optimisation and ROI tracking and provide essential business and operational telemetry through a comprehensive view of key performance indicators (KPIs) such as task success rate and tool-calling accuracy. Finally, this continuous monitoring ensures sustained performance and auditability, allowing organisations to detect model degradation over time and maintain a verifiable audit trail that is essential for meeting modern regulatory compliance standards.

Securing your AI future

Agentic AI represents a significant frontier for enterprise innovation. This blog highlights that successful adoption depends on a strategic, proactive validation approach; a shift from reactive, pre-deployment checks to continuous, integrated practices embedded within business and technical operations.

The inherent complexities of Agentic AI autonomy and non-deterministic behaviour demand a holistic framework encompassing governance, performance, accuracy, safety, and ethical consideration. By integrating validation as a continuous practice, businesses can master autonomous systems and achieve market leadership in the age of Agentic AI.

To navigate this complex landscape, you don't have to tackle the challenge of Agentic AI alone. Deloitte provides the expertise required for your Agentic AI journey. Whether you need our specialist support to design and build your Agentic AI systems, or expert guidance on how to embed rigorous validation within your own operations, we are perfectly positioned to help. Deloitte's comprehensive AI Assurance services provide the independent oversight and validation for your Agentic AI systems, ensuring they are compliant and risk-mitigated, and enabling you to confidently accelerate your path to true market dominance.

Get in touch

Richard Tedder
Partner
AI Assurance
rtedder@deloitte.co.uk

Avtar Benning
Director
AI Assurance
abenning@deloitte.co.uk

Ala Alfakara
Associate Director
AI Assurance
aalfakara@deloitte.co.uk

Bibliography

1 L. Macvittie, "AI Agents vs. Agentic AI: Understanding the Difference," F5, 2025. [Online]. [Accessed 28 10 2025].

2 A. D. Teaganne Finn, "Agentic AI vs. generative AI," IBM, 2025. [Online]. [Accessed 28 10 2025].

3 C. Wang, "AI Agents in LangGraph," Deeplearning.ai, 2025. [Online]. [Accessed 9 2025].

4 L. Yue, S. Kit, Q. Lu, Z. Liming, D. Zhao, X. Xu, S. Harrer and J. Whittle, "Agent Design Patter Catalogue: A Collection of Architectural Patterns for Foundation Model Based Agents," arXiv, 2024.

5 Siemens, "Siemens introduces AI agents for industrial automation," Siemens, 2025. [Online].

6 Salesforce, "Meet Einstein Service Agent: Salesforce's Autonomous AI Agent to Revolutionize Chatbot Experiences," Salesforce, 2024. [Online].

7 JPMorgans, "Payments Unbound: AI Agents," JPMorgans, 2024. [Online].

8 S. Uspenskyi, "Are AI Agents Safe? Potential Risks and Challenges," Springs, 2025. [Online].

9 A. Jain, "Common Challenges and Strategies in AI Agent Development," Oyelabs, 2025. [Online].

10 R. S, "Agentic Drift: Keeping AI Aligned, Reliable, and ROI-Driven," Medium, 2025. [Online].

11 C. Giovine, "Building AI Trust: The key role of Explainability," 26 November 2024. [Online].

12 D. Leslie, "Explaining Decisions Made with AI," SSRN, 2022.

13 M. Kumar, "Transparency and Accountability in Explainable AI: Best Practices," Springer Nature Link, pp. 127-164, 2024.

14 Shetty, "Analyzing AI regulation through literature and current trends," Journal of Open Innovation, 2025.

15 A. Gikay, "Risks, innovation, and adaptability in the UK's incrementalism versus the European Union's comprehensive artificial intelligence regulation," International Journal of Law and Information Technology, 2024.

16 C. Cancela-Outeda, "The EU's AI act: A framework for collaborative governance," Internet of Things, 2024.

17 J. Mokander, "Auditing of AI: Legal, Ethical and Technical Approaches," Digital Society, 2023.

18 Papagiannidis, "Responsible artificial intelligence governance: A review and research framework," Strategic Information Systems, 2025.

19 Leikas, "Ethical Framework for Designing Autonomous Intelligent Systems," Journal of Open Innovation: Technology, Market, and Complexity, 2019.

20 Elliot, "AI Technologies, Privacy, and Security," Frontiers in AI, 2022.

21 Mudry, "The Hidden Dangers of Browsing AI Agents," arxiv, 2025.

22 Piccialli, "AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0," Expert Systems with Applications, 2025.

23 L. Belenguer, "AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry," AI Ethics, 2022.

24 R. González-Sendino, "Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making," Future Generation Computer Systems, 2024.

25 "How to Test AI Agents Effectively," 12 9 2025. [Online].

26 "Agent Evaluation in 2025: Complete Guide," 7 8 2025. [Online].

27 "Agent Evaluation," 22 07 2025. [Online].

28 N. D. M. Y. M. S. Y. Z. Raphael Shu, "Towards effective genai multi-agent collaboration: Design and Evaluation for Enterprise Applications," AWS Bedrock, 2024.

29 "How to Test AI Agents + Metrics for Evaluation," 12 9 2025. [Online].

30 T. H. Y. Z. B. A. Jiarui Lu, "ToolSandBox: A Stateful, Conversational, Interactive Evaluation," Apple, 2025.

31 T. Hall, "Ensuring AI Accuracy in Agentic Analytics," Tableau, 2025. [Online].

32 G. S. D. S. K. G. Luca Gioacchini, "A Modular Benchmark Framework to Measure Progress and Improve LLM Agents," NEC Laboratories, Europe, 2024.

33 UiPath, "Agentic AI," 2025. [Online].

34 A. Jain, "Agentic AI Evaluation: Ensuring Reliability and Performance," techahead, 2025. [Online].

35 M. C.,. R. R. Lareina Yee, "One year of agentic AI: Six lessons from the people doing the work," QuantumBlack AI by McKinsey, 2025. [Online].

36 M. S.-S. Cole Stryker, "What is AI agent evaluation?," 7 8 2025. [Online]

37 Galileo.ai, "galileo.ai," galileo.ai, [Online]. [Accessed Jan 2026].

38 LangChain, "LangChain," [Online]. [Accessed Jan 2026].

39 A. AI, Arize AI, [Online]. [Accessed 01 2026].

Deloitte.

Designed and produced by Creative Studio at Deloitte. J60459