Deloitte.

Together makes progress

Trusted & Secure AI (TSAI)

Agentic AI Application Risk Management

The TSAI Agentic AI Implementation Guidebook offers a structured framework to enhance the safety, security, reliability, and robustness of agentic AI deployments. It provides practical guidance for assessing, addressing, testing, and monitoring AI risks, with a focus on the following key areas: privacy, transparency and explainability, fairness, responsibility, accountability, robustness, and security.

A step-by-step process to embedding trust and security into agentic AI application deployments

01

Assess agentic AI risk profile

Conduct systematic evaluations of agentic AI deployments to **proactively identify risks** related to safety, privacy, and security. Provide targeted recommendations for controls that enable organizations to effectively identify, mitigate, and continuously monitor these risks across the entire AI life cycle.

02

Implement controls to safeguard AI systems

Embed controls throughout every stage of the agentic AI application life cycle, confirming alignment with US government AI regulations, Deloitte GPS AI guidelines, established frameworks such as NIST AI RMF and OWASP, as well as the ethical standards relevant to specific use cases. **Enhance transparency** of agent roles and orchestration workflows, and proactively prevent unauthorized access and unintended outcomes.

03

Conduct quality and risk testing

Perform end-to-end testing of agentic AI applications by assessing risk and quality against defined risk categories and performance metrics. **Verify** that agent reasoning, responses, and behaviors meet established standards and guidelines for accuracy, explainability, and safety.

04

Enable trusted and secure AI operations

Use logging and observability tools to **monitor and evaluate** agentic AI trajectories, memory, and tool usage to identify risks, anomalous behavior, and performance issues. **Continuously refine security and operational controls** in response to emerging threats, system changes, and lessons learned from incidents and audits.

Trusted and Secure AI. **AI built with confidence.**

Deloitte's TSAI Agentic AI Implementation Guidebook provides a tailored risk management approach for agentic AI applications to help unlock Generative AI (GenAI) at scale. From inception through implementation, these seven capabilities—supported by tools and accelerators—offer an effective way to identify, mitigate, and manage risks while improving agentic AI application performance.

Risk assessment

Evaluate LLM and agentic AI application risk to identify effective controls and testing needed to mitigate and manage AI risk.

Capability: Streamlined risk assessment survey designed to target unique use case risks. Customized risk assessment scorecard presenting risk profile and tailored recommendations for risk mitigation strategies.

Application & agent risk management

Implement audits and controls for agentic AI applications at both the individual LLM and agent levels through guardian agents that monitor and intervene in real time to help prevent harmful or sensitive information from being ingested or disclosed, and to curb unintended agent behaviors.

Capability: Oversight mechanisms for agentic AI applications, including deploying guardian agents with authority to monitor, intervene, escalate, or halt actions that violate ethical or policy standards; defining escalation procedures and incident intervention steps with clear owners for oversight and approval; and logging agent decisionsand actions.

Risk acceptance testing

Conduct end-to-end testing and evaluation of safety, privacy, security, compliance, reliability, bias/fairness, and operational risk associated with agentic AI applications.

Capability: Automated prompt testing designed to identify risks and vulnerabilities associated with agentic AI, helping to evaluate the effectiveness of application risk management based on application outputs.

Output quality testing

Measure the completeness, accuracy, helpfulness, relevance, and groundedness of AI-generated outputs to validate autonomous, goal-oriented systems output accuracy, safety, and reliability in alignment with team goals.

Capability: Iterative automated testing designed to evaluate agentic AI performance at three levels—tool/unit tests, agent integration tests, end-to-end simulations (i.e. trajectory testing). A mix of methodologies such as human-in-the-loop, LLM-as-a-judge, and automated checks and scoring are involved for metrics generation.

Monitoring & reporting

Deep visibility into complex agentic workflows, allowing developers to inspect decision paths, tool invocations, and intermediate outputs, and improving transparency into agentic AI operations and performance.

Capability: Specialized, end-to-end observability for multistep AI tasks, reasoning chains, and memory and tool usage, which is summarized as executive scorecards for system and tool performance, risks palettes, and tracing explorer.

Continuous testing

Ongoing monitoring and testing of AI system vulnerabilities and performance—across both the agent and application layers—beyond deployment, by periodically conducting penetration tests against emerging AI threats.

Capability: Periodic reviews and curated threat-simulation test cases aligned to new AI guidelines and emerging vulnerabilities validate the effectiveness of existing controls and surface new risks and performance improvement opportunities.

Get in touch

Joe Conti
Trusted & Secure AI Leader
Deloitte & Touche LLP
joconti@deloitte.com

Mike Ingenito
Trusted & Secure AI Leader
Deloitte & Touche LLP
mingenito@deloitte.com

Jason M. Richards
Trusted & Secure AI Leader
Deloitte & Touche LLP
jasrichards@deloitte.com

Copyright © 2026 Deloitte Development LLC. All rights reserved.