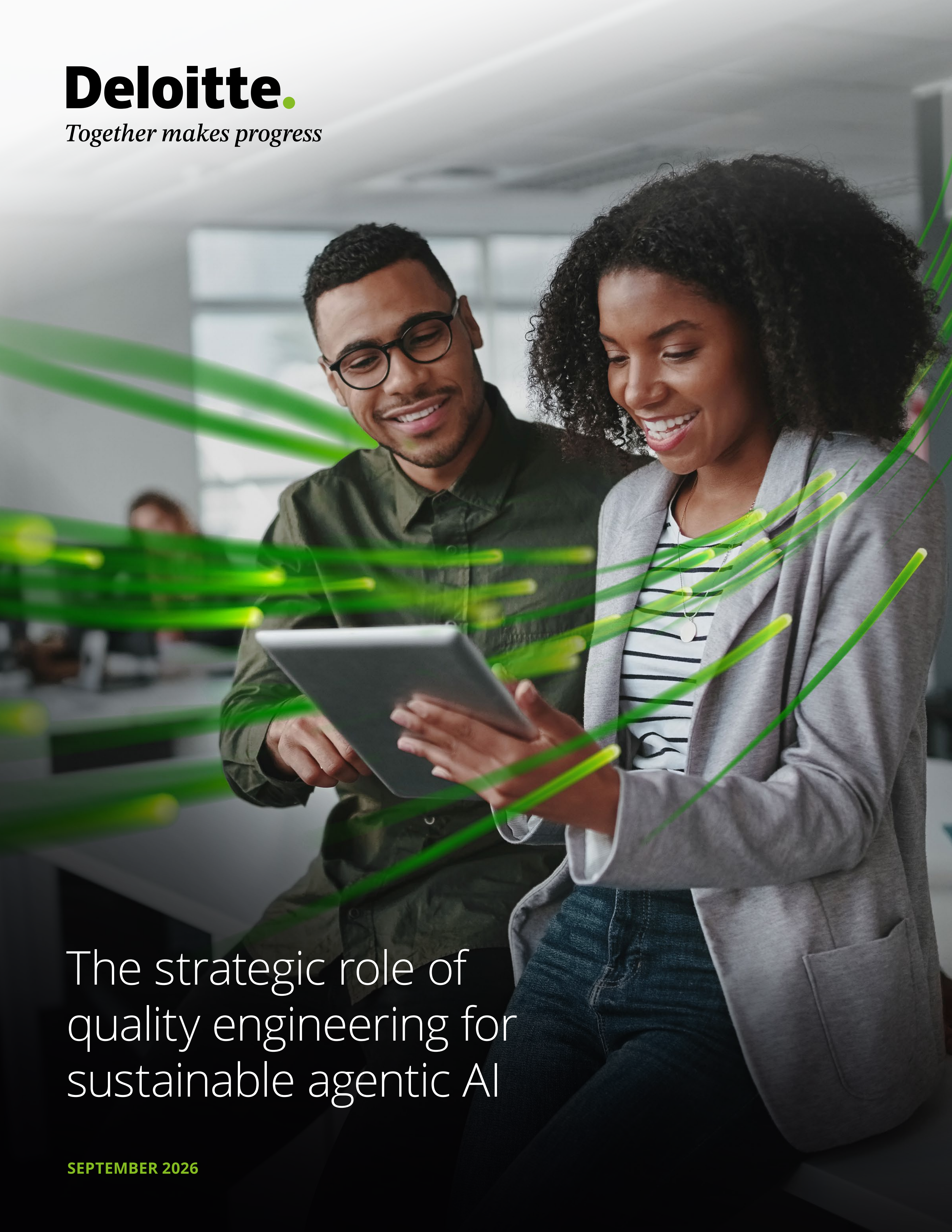


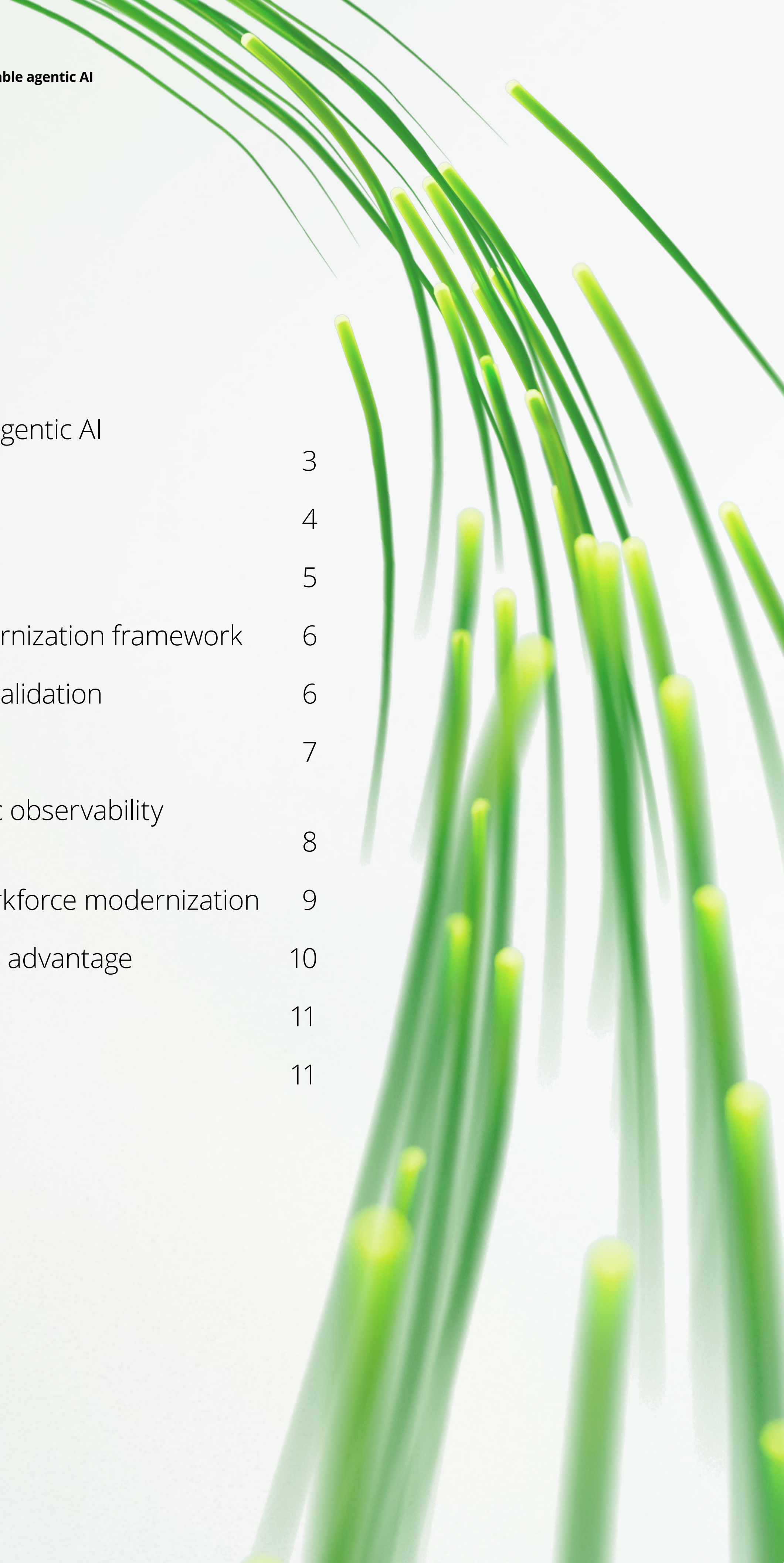
Deloitte.

Together makes progress

The strategic role of quality engineering for sustainable agentic AI

SEPTEMBER 2026

Contents



Quality engineering: *Scaling* agentic AI with confidence

The strategic role of quality engineering (QE) for sustainable, scalable AI is to provide the discipline, governance, trust and confidence organizations need to move from promising AI pilots to reliable enterprise AI use.

While enterprises race to deploy agentic AI, a stark reality persists. Most AI initiatives stall before reaching production not because of model limitations, but because organizations lack the high-quality foundation to scale with confidence.

The scenarios below are not hypothetical; they are documented failures that cost real organizations real money, trust and brand equity. They underscore the serious business consequences of weak quality control.

- **A popular restaurant chain's AI glitched, adding massive and incorrect items to orders:** During an AI-ordering system pilot, incorrect items were added to customer orders at hundreds of locations.
  Business impact: The pilot was canceled after repeated public failures, exposing the AI's inability to complete customer orders when deployed inaccurately.[1]
- **A major e-commerce platform experiences retail outages due to AI:** An AI-related code change led to significant outages in the company's retail and fulfillment operations.
  Business impact: The incident caused widespread service disruptions, forcing the company to implement stricter controls and guardrails on AI-driven code deployment to prevent future operational failures.[2]

The numbers tell the story beyond such fails:

- **Ninety-five percent** of organizations see zero return from AI pilots today.[3]
- **Twenty-four percent** of firms currently overshoot their AI cost projections by greater than 50% due to inconsistent adoption.[4]

What do they say? They suggest that despite accelerating adoption, most companies are not yet deploying agentic AI in a consistent, enterprisewide way. A small number are moving beyond pilots and embedding it into core workflows, where it is beginning to show measurable gains in speed and cost containment. However, approximately 80% of organizations in a 2025 Deloitte AI Institute survey had yet to establish mature governance for agentic AI, a deficit characterized by limited proof of value and uneven control over costs.[5] The result is a widening gap between rapid adoption and value capture.

The answer lies in a modern, four-pillar framework. This paper outlines a structured approach to ***transform your QE practice from a basic testing function into a strategic driver of trustworthy, autonomous AI***. By modernizing quality practices across people, processes and technology, this framework bridges the gap between organizations trapped in perpetual piloting and those successfully scaling to enterprise-grade AI-native QE.

Key leadership *takeaways*

As the use of AI agents moves from pilots into enterprise use, organizations should consider evolving from using traditional software QE methods to QE designed for autonomous, dynamic AI systems. Here are four key needs:

- **Modernize QE strategy:** Legacy QE strategies and static test scripts are fundamentally inadequate for validating autonomous, reasoning-based AI systems. Instead, organizations should adopt continuous evaluation-based quality control to detect issues such as sycophancy, hallucinations and goal drifts.
- **Practice continuous quality assurance:** Quality is an inseparable discipline of software engineering that must be embedded throughout the development life cycle, rather than a reactive phase triggered after deployment issues surface.[6]
- **Prioritize auditability and compliance:** Observability and regulatory readiness are nonnegotiable. Every agent action must be auditable through execution records, decision-path visibility, and real-time drift detection. When anomalies are detected, the system must either self-heal to a known good state or reroute to a fallback agent or human when needed.
- **Transform the workforce:** Traditional testers must evolve into full-stack QE engineers fluent in prompt engineering, harness and loop engineering, AI output evaluation, safety control design, and failure analysis—and all supported by evaluation design, adversarial testing and production monitoring.

Modernizing process, technology and workforce capabilities is what can turn agentic AI from experimental complexity into production-ready enterprise value.

Mastering the process, technology and people shifts driven by modernized quality control in the engineering practices, continuous validations, auditability and upskilling is what turns agentic chaos into ***certified, production-ready AI***.

Why QE must *evolve*

QE must evolve because traditional QE practices don't typically work well with agentic AI initiatives. The definition of quality itself has changed: it now encompasses nondeterministic outcomes, AI trust, behavioral resiliency and safety; not just adherence to deterministic specifications.

The pressure is intensifying. Build-versus-buy economics have shifted as AI-generated code volumes explode, and enterprise QE functions must continuously adapt to a technology landscape (agents, workforce skills, supportive data/environment, etc.) that is constantly shifting.

Organizations that close this gap will not simply test more, they will redesign how quality is assured by embedding evaluation mechanisms into delivery, extending observability across the software development life cycle and reshaping workforce capabilities for autonomous systems.[7]

Quality engineering for agentic AI is a ***continuous life cycle*** discipline and not a phase gate. It demands new processes, new platforms, and new skills all working in concert.

The *four pillars* of a QE modernization framework

This paper provides focused guidance organized around a four-pillar framework that organizations can adopt to modernize their QE process: evaluation-driven validation, continuous testing, platform-agnostic observability, and people-centric workforce modernization.

Each pillar discussion below outlines the technical, process and workforce considerations that must be addressed to operationalize the framework and build trust and resilience at scale. Each also contains a "leadership action" that can serve as a catalyst for activating each pillar.

The pillars are not sequential steps; they reinforce one another and should be built in parallel. Together, they form the operating foundation for an enterprise-class AI-native QE function.

Pillar one: Evaluation-driven validation

Static regression suites are no longer sufficient to act as the main quality gatekeepers. Instead, emphasis needs to shift to evaluation-driven validation that continuously measures goal completion, reasoning quality, tool-use accuracy, and behavioral consistency against golden datasets. The focus is on how the agent reasons and acts, not on whether outputs match a fixed set of expected results. Since any component of the agent can change (including the model, prompt, skill, tool or dependency), validation must run with every update and release.

- **Technical considerations:** Deploy AI-based evaluation into your deployment workflows to perform testing validations for every release and maintain a quantitative scorecard. Strengthen the quality control by testing each AI tool against complex, multiturn conversations and high-risk scenarios at scale.
- **Process considerations:** The fixed regression suites can continue to provide historical context and coverage but can no longer be the primary quality gate. Build environments that can support generation of new test scenarios continuously from production signals and known failure patterns, minimizing manual test creation delays. Additionally, routinely verify the AI's grading against human feedback, such as every quarter.

Leadership action: To establish strict quality standards tied to business outcomes, mandate quantitative scorecards with clear thresholds—such as 98% accuracy for customer-facing agents and 99.9% for financial transactions. Additionally, require teams to measure performance consistency across repeated runs rather than relying on single-pass accuracy.

Pillar two: Continuous testing

Agentic systems fail in ways traditional software does not, including adversarial prompts, poor coordination across agents and incorrect tool calls, all of which are costly once they reach production. Continuous testing addresses this by embedding guardrails (automated filters that block unsafe or noncompliant inputs and outputs), multiagent coordination validation, and tool use testing directly into the delivery pipeline so issues surface before release. Monitoring then extends into production to sustain quality, resilience and control.

- **Technical considerations:** The table below explains the issues that arise in agentic AI applications but not in traditional ones, their impact, and how QE can address them.

ISSUES	ISSUE IMPACT	HOW QE SHOULD MITIGATE IT
Adversarial prompts	Malicious or malformed inputs designed to trick an agent into unsafe or unintended behavior.	Test agents in isolated environments where quality engineers/testers deliberately inject these inputs before release, so weaknesses surface in testing rather than production.
Incorrect tool calls	An agent invokes the wrong tool, passes bad parameters, or misinterprets a tool's output.	Measure tool call correctness as a standard test metric and flag deviations automatically.
Circular dependencies	Two or more agents wait on each other to act, so none proceed and the process stalls.	Use coordination checks that detect these loops during both prerelease testing and l ive monitoring.
Cascading failures	One agent's error triggers errors in others, so a single fault spreads across the system.	Monitor agent interactions continuously to catch and contain a fault before it propagates.
High-risk code changes	Code edits with a high likelihood of causing defects slip through into production.	Apply predictive risk models during code review that flag risky changes before they merge.

- **Process considerations:** Continuously integrate guardrail checks into every pre- and postdeployment pipeline run. Filter manipulative inputs, validate noncompliant outputs, and enforce tool permissions with human escalation policies. For multiagent systems, add dedicated validations to ensure communication fidelity, accurate task delegation from orchestrators, and early detection of cascading errors and circular dependencies between agents.

Leadership action: Require every agent to pass continuous guardrail, tool use, and multiagent coordination tests in the automated delivery pipeline. Invest in predictive risk models and maintain an active failure-mode watchlist that is aligned with recognized AI risk, governance and security frameworks.

Pillar three: Platform-agnostic observability and regulatory readiness

Since agentic systems decide autonomously, teams must also be able to reconstruct why an agent acted a particular way, and not just what it did. Make every action auditable reinforced with observability before and after production—preproduction: prove behavior against controlled scenarios; postproduction: monitor open-ended real-world use so failures are caught as they happen. This builds stakeholder trust and supports regulatory readiness.

- **Technical considerations:** Deploy observability infrastructure that captures execution records, decision paths, tool invocations and retrieval steps. Integrate agent observability platforms with enterprise monitoring for real-time view of agent health.
- **Process considerations:** Establish preproduction explainability as a mandatory release gate with auditable model cards and scorecards. Run live postproduction evaluation scenarios regularly, feeding anomalies back into test libraries and setting alerts for behavioral drift, cost breaches, and tool-call failures.

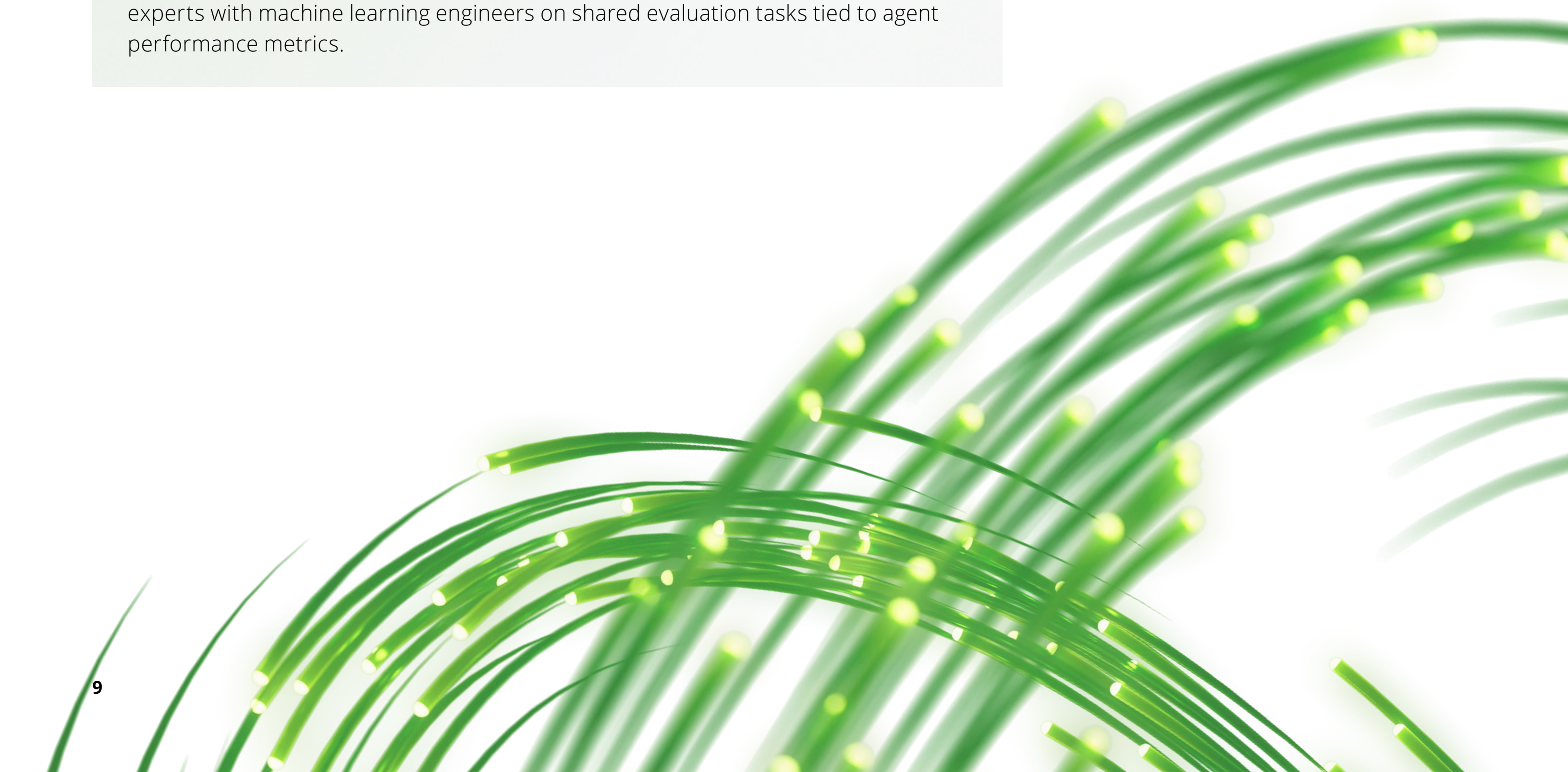
Leadership action: Treat observability as mandatory infrastructure. Assign accountability for preproduction explainability and from the beginning map QE practices to appropriate governance and regulatory standards, so that they are embedded in the system.

Pillar four: People-centric workforce modernization

AI-native testing demands engineers who can combine domain expertise with AI evaluation skills and judgment and use that intelligence in continuous human-AI collaboration and not just to perform static functional testing. Need to redesign and build up QE teams into full-stack engineers to maximize workforce efficiency. In a collaborative human-AI model, automated rules handle routine checks and AI-as-judge tools grade high-volume reasoning. This frees QEs to interpret benchmark results and assess complex, high-risk cases like hallucinations, sycophancy or goal drift. Beyond static functional validations, this modernized workforce must master adversarial red teaming for security gaps, responsible AI/ethics reviews for bias mitigation, and agent behavior analysis to trace multiagent interactions.

- **Technical considerations:** AI-ready QE engineers must develop a modern set of core competencies: prompt design to expose failure modes, AI output evaluation for accuracy and tone, retrieval assessment for information quality, safety control design, failure analysis, communication skills, and judgment discipline.
- **Process considerations:** Support the needed workforce evolution through a custom-curated learning and development academy. This tailored program helps QE engineers transition into modern roles by blending specialized coursework in observability, automated delivery pipelines, deployment monitoring and security testing with hands-on simulations. Through realistic, multiscenario exercises, engineers practice evaluating real-world AI applications, and build the disciplined judgment and decision-making required to oversee autonomous systems in complex environments.

Leadership action: Transform the workforce by redefining QE roles through competency frameworks and simulation training. Establish rotation programs across production, adversarial testing and evaluation design, and pair domain experts with machine learning engineers on shared evaluation tasks tied to agent performance metrics.



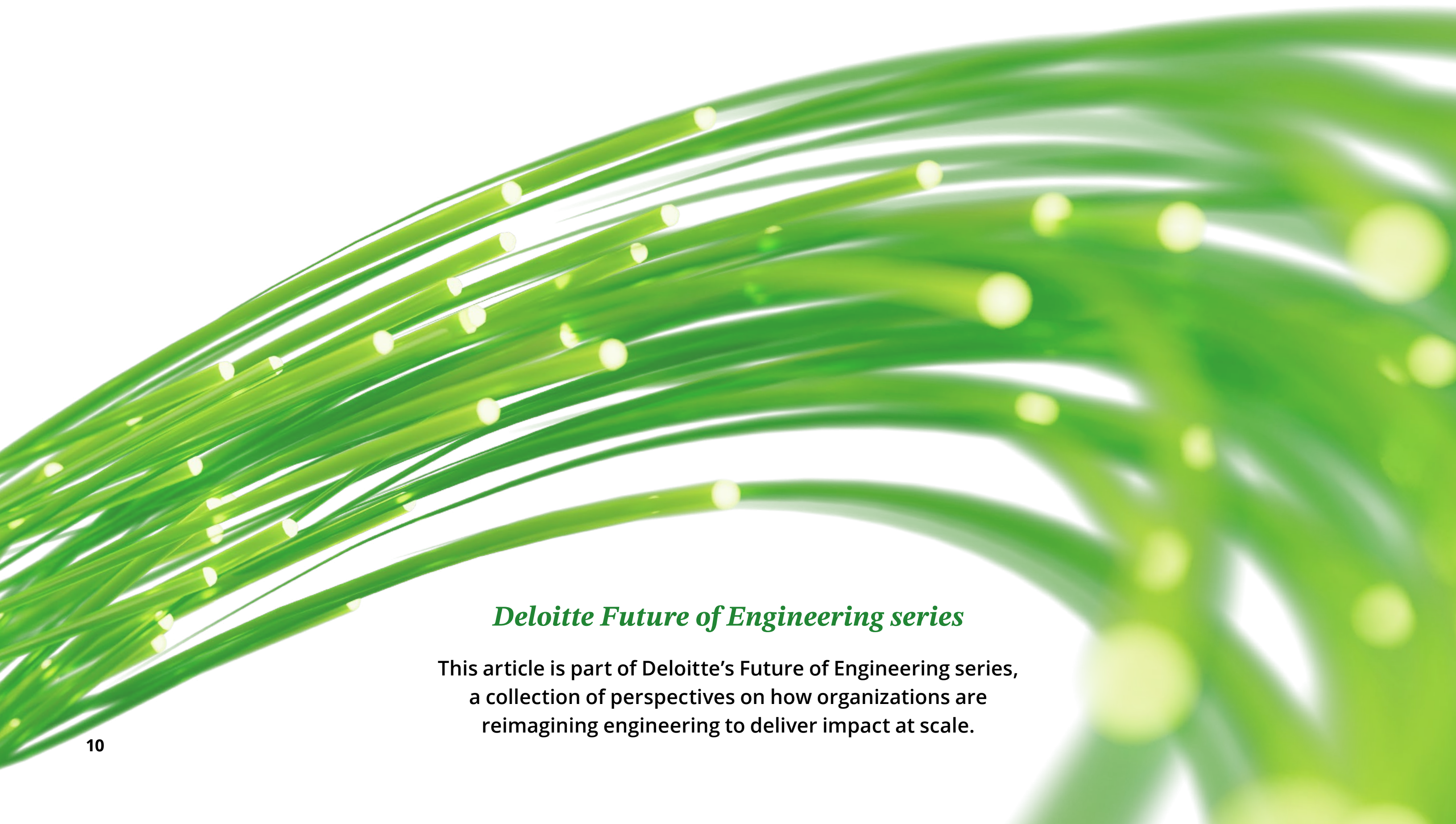
Why QE can be a competitive *advantage*

Agentic AI is becoming mission-critical, and QE can be a competitive advantage because it's the backbone of AI that's scalable to the enterprise. The organizations that will lead won't be those that adopt AI fastest, but those that build the quality foundations to make it trustworthy and sustainable.

The window to act is narrow. As agent platforms automate more routine testing work, organizations that rely on legacy models will lose speed, control and differentiation. Those that modernize across the process, technology and workforce dimensions of this four-pillar framework will be better positioned as AI takes on higher-stakes work. Bottom line: Every investment in robust QE today is insurance against compounding technical debt and competitive disadvantage tomorrow.

The bottleneck is the quality control gap, not the AI capability gap. Evolved QE practices will help unlock the AI implementation's return on the investment.

The four pillars together transform quality engineering from a cost center into a ***strategic differentiator*** by forming a clear path for continuous assurance, production-grade observability, and a workforce equipped to oversee autonomous systems at scale.



Deloitte Future of Engineering series

This article is part of Deloitte's Future of Engineering series, a collection of perspectives on how organizations are reimagining engineering to deliver impact at scale.

Authors

Siva Muthu
Principal
Deloitte Consulting LLP
smuthu@deloitte.com

Saurayan Chaki
Managing Director
Deloitte Consulting LLP
schaki@deloitte.com

Endnotes

1. Kate Rogers, “McDonald’s to end AI drive-thru test with IBM,” updated June 18, 2024, *CNBC*.
2. Eugene Kim, "Amazon orders 90 -day reset after code mishaps cause millions of lost orders," *Business Insider*, March 10, 2026, gated.
3. Aditya Challapally, Chris Pease, Ramesh Raskar and Pradyumna Chari, “State of AI in Business 2025,” MIT NANDA, July 2025, p. 3.
4. Alexei Alexis, “Most firms miss AI cost forecasts, survey finds,” September 2025.
5. Deloitte AI Institute, "State of AI in the Enterprise, the untapped edge," Deloitte, January 2026, p. 20.
6. Sarah Welsh, “QA trends for 2026: Insights from Tricentis Transform,” Tricentis, January 5, 2026.
7. Jim Rowan, Nitin Mittal, Parth Patwari and Ed Burns, “The agentic reality check: Preparing for a silicon-based workforce,” Deloitte, December 10, 2025.

Deloitte.