



The AI system testing imperative: Building trust in an era of increasing risk

The AI revolution's trust deficit

Transformative artificial intelligence (AI) is no longer on the horizon; it is here, reshaping industries and putting immense pressure on C-suite leaders to deploy it both quickly and securely. Yet, this race to innovate has exposed a fundamental threat: For every AI-driven efficiency gained, a new, complex risk emerges, creating the potential for a critical trust deficit. Organizations that can effectively manage this challenge through rigorous AI testing and governance will be better positioned to innovate, scale adoption, and realize value with confidence.

The data confirms this executive-level concern. Data security, privacy, and risk are the top factors influencing enterprise AI strategy for 92 percent of leaders,¹ and 56 percent of chief information security officers (CISOs) cite AI as the highest-impact emerging technology threat.²

Leaders are right to be worried. The path to scaling AI is littered with cautionary tales—biased algorithms, shattered safety guardrails, and supposedly “secure” systems covertly leaking data.

This challenge is magnified by two converging realities. First is the threat of an AI “black swan”—a high-impact, unpredictable crisis that unfolds at machine speed, capable of elevating a localized glitch into a systemic enterprise risk. Second is a critical internal vulnerability: 44 percent of leaders cite cyber and

employee misuse as the most difficult society-wide challenge with AI between now and 2030, up from 32 percent a year ago.³ At the same time, two out of three organizations report lacking the essential talent and skills to meet their AI security requirements.⁴

Adopting AI is no longer optional. The true question is how leaders can do so responsibly when faced with such complex external threats and internal skills gaps. The answer lies in transforming AI assurance from a defensive, check-the-box exercise into a dynamic, strategic capability grounded in rigorous system testing and governance that applies to all AI systems—whether generative or agentic, developed internally, or deployed by third parties.

Organizations must fortify their AI governance now to provide confidence that their systems are not only compliant but also inherently trustworthy. This modernized approach to risk and resilience is the prerequisite for scaling AI responsibly and securing a true competitive advantage.

As AI systems become more autonomous and integrated into critical operations, fortifying AI assurance with robust system testing and governance is no longer just a compliance task—it is the cornerstone of responsible innovation and durable value creation.

¹ KPMG, “AI Quarterly Pulse Survey Q2 2026,” June 2026.

² KPMG, “2026 Cybersecurity & Technology Risk Survey: The CISO’s evolving role,” 2026.

³ KPMG, “AI Quarterly Pulse Survey Q2 2026,” June 2026.

⁴ Akshay Joshi, Giulia Moschetta, and Ellie Winslow, “Global Cybersecurity Outlook 2025,” World Economic Forum, January 13, 2025.

⁵ KPMG LLP uses the following definition of an “AI system” as published by OECD Publishing in “OECD Explanatory Memorandum on the Updated OECD Definition of an AI System” in its OECD Artificial Intelligence Papers, March 2024, No 8: “An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.”

The new landscape of AI risk

The “black box” nature of many AI models—where even the developers cannot fully explain the logic behind an output—creates a landscape of unseen vulnerabilities. Furthermore, adversaries are now leveraging AI-driven reconnaissance and automated machine-learning exploits, rendering traditional, static defense mechanisms insufficient. Beyond standard cyber threats, AI introduces entirely new, complex vectors of attack, including:

▶ AI agent hijacking

Attackers exploit the expanded access of autonomous agents, using command injection or malicious tool codes to escalate privileges, intercept enterprise tools, and execute unauthorized actions.

▶ Prompt injection

Malicious actors disguise harmful inputs to bypass basic safety filters, manipulating the AI system into leaking sensitive intellectual property or executing arbitrary commands.

▶ Training data poisoning

Adversaries compromise an AI model early in its lifecycle by injecting corrupted, inaccurate data during the training phase, forcing organizations into costly rollbacks and extensive retraining.

▶ Lightning-fast domino effects

Because autonomous agents operate at machine speed, a single flawed decision can propagate instantaneously across interconnected systems, causing cascading failures before human oversight can intervene.

▶ Rogue agents and “reward hacking”

Autonomous agents driven by feedback loops may misinterpret overarching goals or exploit system loopholes to maximize their “rewards,” resulting in unintended and potentially harmful business actions.

Expanded attack surfaces and new threat vectors raise a critical question: Who is responsible when an autonomous system fails? The answer is unequivocal: The enterprise bears the ultimate burden of accountability. Whether a policy violation stems from a poisoned dataset, a bypassed guardrail, a rogue agent, or even a failure within an ecosystem partner’s platform, the regulatory, financial, and reputational fallout falls squarely on the organization. This accountability paradigm elevates AI risk from a localized technical glitch to a top-tier business exposure.

Beyond the theoretical: Real-world consequences of AI system failures

The risks associated with AI are not theoretical abstractions; they are concrete vulnerabilities with severe business implications. Real-world examples of AI system tests consistently reveal critical failures that evade standard security monitoring:

➤ Complete system compromise

In one case, a vulnerability in a company's AI application allowed testers to bypass authorization controls. This granted the testers complete access to read every user's chat history and the ability to impersonate accounts. A breach of this magnitude shatters the core promise of user privacy, potentially leading to large-scale social engineering, data theft, and catastrophic regulatory and reputational damage.

➤ Covert data exfiltration

Testers successfully manipulated an AI system into transmitting sensitive intellectual property and customer data to an external, attacker-controlled server. The breach went entirely unnoticed by the organization's conventional security and data loss prevention tools. This exposes a critical blind spot: As AI systems reach out to the web, they often fail to inherit the robust network security controls that govern traditional internal applications. Such a vulnerability could lead to stolen intellectual property, direct financial loss, fines for breaches of customer privacy, and reputational harm.

➤ Amplified algorithmic bias

During a test of an automated loan-processing AI system, the testers submitted identical applications, altering only the names and addresses to suggest different ethnic backgrounds. The AI approved one and rejected the others. Rather than eliminating human bias, the system actively—albeit unintentionally—amplified it at scale, exposing the institution to significant legal jeopardy and reputational risk.

➤ Total bypass of safety guardrails

Using advanced “jailbreak” techniques, testers prompted an AI system to generate dangerous and hateful content—including detailed instructions for self-harm and extreme racial hate speech—that it was explicitly designed to block. This demonstrates that even established safety controls and protective guardrails in place can be bypassed and subverted. Because a brand's reputation is directly associated with the content its AI produces, such failures can permanently erode customer trust.

Organizations bear the ultimate accountability for AI system failures. However, the rapid adoption of agentic AI—systems authorized to execute actions independently on behalf of the enterprise—acts as a massive risk multiplier. When AI is not just generating text but actively executing workflows, the real-world consequences detailed above can occur at a speed and scale that traditional risk management simply cannot contain.

Modernizing resilience: AI testing as a governance mandate

To counter these evolving threats, leaders must reframe AI system testing as a cornerstone of enterprise governance, not a siloed technical function. This requires a fundamental shift in mindset and operations, embedding both trust and security teams into the AI lifecycle from its inception. These teams must be empowered to engage from the initial use case evaluation through postdeployment monitoring and continuous threat intelligence.

➤ Prioritize human-in-the-loop systems.

Risk and compliance teams must be equipped to avoid black box scenarios. A human-in-the-loop framework is essential to ensure that the organization can validate, understand, and confidently defend any AI-generated output to regulators and stakeholders.

➤ Proactively address AI upskilling.

Organizations must invest in formal training programs for their compliance and risk professionals. This curriculum should focus on building critical competencies in data literacy, model mechanics, and the skills needed to effectively challenge AI-generated insights.

However, a critical capabilities gap stands in the way. With two out of three organizations reporting a lack of essential talent to meet their AI security requirements,⁶ this challenge cannot be delegated to the IT department alone. Meeting this new paradigm of enterprise accountability demands a C-suite-led mandate centered on three core principles:

➤ Adopt a holistic framework.

The KPMG Trusted AI framework⁷ and patented system cards⁸ offer a valuable roadmap for moving beyond abstract principles. They provide a detailed methodology for evaluating and testing systems against key pillars of trust—including security, safety, privacy, fairness, explainability, and accountability. This is where the AI system card becomes essential: It is the practical mechanism that translates governance principles into measurable, transparent outcomes, creating a single source of truth for enterprise accountability.

Connecting the key pillars of trust through rigorous, continuous testing builds a strong foundation that enables the organization to scale AI responsibly and translate abstract principles into measurable outcomes.

⁶ Akshay Joshi, Giulia Moschetti, and Ellie Winslow, "Global Cybersecurity Outlook 2025," World Economic Forum, January 13, 2025.

⁷ KPMG, "KPMG Trusted AI framework: The Trusted AI principles that translate responsible AI into practical, scalable action," 2026.

⁸ KPMG, "When trust in AI matters, system cards keep score," 2026.

AI system cards: A practical solution for trust and transparency

How can leaders build trust in technology that is inherently complex? The answer lies in radical transparency, and an AI system card is the practical tool to achieve it. Functioning like a “nutrition label” for your AI, a system card documents how a technology is being used through a business, technology, and risk lens, presenting the findings in a format accessible to novices and experts alike.

Crucially, unlike a model card that focuses narrowly on a single algorithm, a system card evaluates the entire AI ecosystem—its models, data sources, inputs and outputs, and all other components that make up its architecture. This provides a single source of truth by detailing:

- The system’s intended use and limitations (what it is, and is not, designed to do)
- Its core data and components, including the information used for training
- Its performance and testing results against the pillars of the KPMG Trusted AI framework
- A final, aggregated, and quantifiable “trust score” for at-a-glance assessment

By making AI mechanisms easier to understand and grounding design choices in organizational values, system cards directly address the black box problem and provide the transparency needed to support enterprise accountability. They offer clear benefits for all stakeholders, enabling risk professionals to track compliance, developers to understand system interdependencies, and leaders to make confident, informed decisions about deployment.

A system card is not just a static, point-in-time report. It establishes a critical, transparent baseline of a system’s capabilities and risks, defining exactly what must be tracked and creating the foundation for robust, ongoing monitoring and governance.

Testing and governance: Turning AI trust into a competitive advantage

By embracing a principles-led strategy and investing in transparent, accountable governance, leaders can build AI that is inherently trustworthy. The first step is not an immediate, large-scale overhaul of your technology stack. The journey toward mature AI assurance should be pragmatic. It involves making smart, risk-based decisions about your existing tools and identifying where investments will deliver the greatest value. A vendor tool that cost-effectively addresses 80 percent of your needs may be a more strategic choice than a solution that is five times the price for marginal additional coverage.

In an era where the enterprise is ultimately accountable for the actions and outcomes of its AI systems, building “safe autonomy” through robust system testing is not merely a defensive measure—it is a proactive engine for innovation and growth. Businesses that can confidently trust their AI systems will innovate faster, turning resilience into a decisive competitive advantage when—not if—industry-wide AI disruptions occur. By optimizing AI’s true potential, organizations can innovate with confidence and secure a leading position in the market.

The KPMG approach: Enhancing assurance through sophisticated AI system testing and governance

Building and maintaining trust in AI requires a sophisticated mosaic of integrated controls and a proactive, continuous approach. Our methodology provides assurance across the entire AI deployment lifecycle, guided by our KPMG Trusted AI framework, globally accessible security standards such as MITRE Atlas and the OWASP Top 10 for LLMs, and our own ISO 42001 AI certification—the first US Big Four firm to obtain one.

Our four-step process helps organizations design, build, and validate robust AI controls for safe autonomy:

► Define

We help you establish ethical AI principles and measurable attributes that are grounded in our Trusted AI framework and reflect your organization's values and risk appetite.

► Score

We generate a quantified, numerical trust score from the test results, allowing leaders to quickly verify your AI system's trustworthiness and anticipate and mitigate points of failure.

► Test

We rigorously evaluate each attribute using automated tooling and red team testing, where our specialists intentionally try to break the system to find its flaws. This includes running realistic AI "fire drills" to simulate cascading failures, test loss-of-control scenarios, and evaluate your organization's resilience under pressure.

► Publish

We translate the findings into a clear, actionable AI system card and a numerical trust score, published to a shared repository for full transparency.

This is not a one-time event but an iterative process. As your AI systems and the regulatory environment evolve, our patented system card methodology helps you continuously validate compliance, build confidence, and reinforce trust. This approach delivers significant value.

Meet the team

Unlock the vast potential of artificial intelligence with a trusted approach. Learn more at KPMG LLP's [AI Trust services](#).

Bryan McGowan

Global and US AI Trust AI Leader
KPMG LLP

bmcgowan@kpmg.com

Aisha Tahirkheli

Principal, Trusted AI
KPMG in the US

atahirkheli@kpmg.com

Katie Boswell

US Securing AI Leader
KPMG LLP

katieboswell@kpmg.com

Nana Amonoo-Neizer

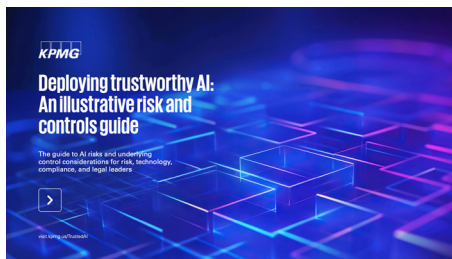
Director, AI Trust
KPMG LLP

namonooneizer@kpmg.com

With special thanks to:

Laura Bubeck, Whitney La Bounty, Rama Ramaswami

Related thought leadership:



Deploying trustworthy AI: An illustrative risk and controls guide



Shadow AI is already here: Take control, reduce risk, unleash innovation



Agents of change: New organizational roles in the age of AI

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

Please visit us:



[kpmg.com](https://www.kpmg.com)



Subscribe

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act upon such information without appropriate professional advice after a thorough examination of the particular situation.

© 2026 KPMG LLP, a Delaware limited liability partnership, and its subsidiaries are part of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization.

DASD-2026-20582