Deloitte

Together makes r

Advanced safeguards are essential to capturing AI value and business growth

Deloitte Trustworthy AI™ Solutions: **Cyber AI Blueprints and Technology Services**

Security, risk, and privacy are among the most serious challenges to businesses adopting artificial intelligence (AI) at an enterprise level.[1] As AI adoption grows, misuse, deepfakes, data poisoning and exfiltration, model inversion, model bias, leakage of intellectual property, and other risks are part of the expanded threat landscape. The autonomy built into emerging agentic AI systems amplifies those threats, while the use of third-party AI resources—no matter how trusted—limits visibility and magnifies exposure.

A 2024 Deloitte study[2] found almost one-third of surveyed organizations said managing risks was a significant barrier to developing AI a six point increase from less than a year earlier. Yet, in an online poll[3] of business leaders, only about one in 10 reported their organization had a comprehensive cyber AI risk management program up and running.

Customer trust in a brand does not necessarily extend to AI systems that the brand uses. Designers, creators, owners, and users need to know that when an organization builds an AI system, they can trust it that it is secure, private, resilient, accurate, and aligned with their interests. That means cyber is much more than vigilance at the entry points and across attack surfaces. To protect and empower AI, cyber should begin at the beginning, as a consideration in a modernized software development life cycle. While it's possible to reverse engineer safeguards later on, it is seldom as effective and typically more costly. Another complication is that the unauthorized use of public or open source AI tools that are popular in the marketplace, but outside an organization's own development and governance, can create a "shadow IT" that offers a side door to threats.

Cyber AI Blueprints and Technology Services is not only about a strong defense. It's also a key to capturing value, namely operational efficiency, time and cost savings, and scale (or speed). Trust is what unlocks AI's potential, enables adoption, and powers consistent outcomes so the business can realize the full benefits of technology transformation. If cyber is not already a key component of your AI strategy, now is the time to shift.

A complex arena of security and value

The relationship between cyber and AI is symbiotic: Cyber protections secure AI resources and help promote operational efficiency, cost savings, time savings, scale, and speed to market. On the other hand, AI can help make cyber protections more effective, scalable, and efficient. Both are valuable, but securing AI is step one. To put it in simple terms, if your agentic AI solution is operational and can either be manipulated or goes rogue of its own accord, you'll be doing bad things faster and more efficiently—which is problematic.

Addressing these needs is ideally part of an AI program from the start, but the ideal isn't always possible. One organization may have yet to begin using AI animated by an enthusiasm that doesn't focus enough on security and privacy. Another might be farther down the path, perhaps with a mix of owned and third party uses, and want to get more value and security by tightening its approach. A more advanced AI organization may still see the opportunity to expand AI into ambitious new areas with the proper guardrails.

Up to date across the board

Cyber protection for AI systems isn't something that an organization can layer on at the end. Building it in from the ground up involves a trusted AI architecture that integrates safeguards throughout development, deployment, and runtime. That doesn't have to mean building a new cyber program from scratch, as many of the risks and many of the countermeasures are related to well-established approaches. It does, however, call for a modernized approach that covers the range of AI development and use:

- **Programmatic AI governance:** An enterprise wide, systematic approach to overseeing the development, deployment, and monitoring of AI within an organization.
- **AI life cycle management:** Tools and processes that help mitigate operational risk and standardize business and data science practices across the AI life cycle.
- **AI testing:** Evaluating and validating AI models for accuracy, reliability, performance, and fairness.
- **Trust operations:** Configuration and deployment of solutions and technology for continuous monitoring of AI interactions, model performance, and adversarial activity.
- **Data quality, privacy, and safety:** Protocols to prepare correct, balanced, unbiased, appropriate data for development, training, and testing.
- **Platform security:** Securing the underlying infrastructure and software platform that hosts AI systems.

Cyber AI Blueprints and Technology Services in practice: Case example

Many organizations across a variety of industries have turned to **Cyber AI Blueprints and Technology Services** to help build or amplify that protection.

In one example, a global publisher that specialized in educational, government, and business content had embraced a transition from print to digital, not only in its products but by providing an AI powered chatbot for customers. The data that informed the chatbot's large language model (LLM) was proprietary, but the safety controls that protected it were limited. The risk of data leakage, harmful content, and malicious action was unacceptably high.

Before launching the tool, the company wanted to bolster security while also streamlining the user experience. This called for a balance: Greater safeguards might add latency to the responses customers sought. The Deloitte team evaluated a range of guardrails, tested and configured the highest priority solutions, recommended custom safety rules and prompts, and reviewed the AI security framework that was already in place using known threats as a benchmark. New, AI specific threats were added to the chatbot's threat library, along with countermeasure controls.

As a result, the publisher saw an improvement in the security and trust of the consumer facing AI chatbot, with a repeatable testing approach in place, assets to evolve the guardrails as needed, and an initial security framework that it could pilot across other areas of the enterprise.

Tools that can bring Cyber AI to life

The **Cyber AI Blueprints and Technology Services** team at Deloitte helps organizations protect and accelerate enterprise AI at scale by pairing business advisers and cyber AI specialists who combine leading technologies with AI fueled automation, offering industry leading solutions in design, integration, and managed services. With these complementary capabilities, we help organizations transform, safeguard, and enable their operations to be future ready.

Deloitte uses an array of proprietary Generative AI tools in combination with third party assets to carry out this mission. Using a secure software development life cycle approach, we help clients speed development; reduce risk; and automate oversight, incident management, and remediation, all through a detailed executive dashboard.

Capabilities that help organizations establish cyber governance of AI systems include:

- **Secure AI data management,** including application and model inventory, data provenance and lineage, analysis and segmentation, labeling, and synthetic data generation.
- **Secure AI application development,** which focuses on architecture, impact and threat assessments, application and model design review, model tuning and hardening, model observability, and model documentation.
- **Secure AI deployment,** including model stress testing and validation, "red teaming," firewalls and guardrails, prompt and model monitoring, model registration, and AI use notification and disclosure.
- **Foundational security capabilities** that range from asset management to data protection, identity and access management, data loss prevention, threat and incident management, physical and supply chain security, vendor risk management, and more.

AI risk is enterprise risk—and managing it properly helps unlock potential

To a greater extent every day, AI is inseparable from the enterprise. But the breadth of use cases, AI capabilities, and organizational maturity levels makes a one size fits all approach to security unrealistic.

However, by understanding the technology, the threats, the enterprise, and the place where they each meet, an organization can support its AI program with need and threat specific safeguards that not only reduce the likelihood and consequences of negative outcomes, but also clear the way for AI to deliver more of its intended value.

The Deloitte difference

At Deloitte, we believe trust is essential to scaling AI with confidence. It must operate on two levels: the system must be designed to perform reliably, and people must feel secure in using it.

That's why we've built an integrated platform for Trustworthy AI one that brings together machine level governance and human centered design. It's engineered to help organizations develop AI systems that are secure, transparent, explainable, and aligned with intended outcomes.

Backed by Deloitte's AI Institute, supported by global research, and informed by deep experience across industries, this platform helps organizations embed trust into AI development from day one transforming it from a reactive concern into a proactive capability for responsible growth and long term value.

Mark Nicholson
Cyber AI Leader
Principal
Deloitte & Touche LLP
manicholson@deloitte.com
+1 917 952 1014

Cyber AI Blueprints and Technology Services

1. Jim Rowan et al., Now decides next: Generating a new future, Deloitte State of Generative AI in the Enterprise Quarter four report, Deloitte, January 2025, p. 14.
2. Ibid.
3. Live participant poll during Deloitte Dbriefs webcast "Enabling the AI fueled business with cyber AI and automation," October 8, 2024.
4. Deloitte, Trustworthy AI™ framework, accessed July 2025.

This article contains general information only and Deloitte is not, by means of this article, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This article is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional adviser. Deloitte shall not be responsible for any loss sustained by any person who relies on this article.

As used in this article, "Deloitte" means Deloitte & Touche LLP, a subsidiary of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.