Deloitte.

Together makes progress

Trust begins at the core

Deloitte Trustworthy AI™ Solutions: Model Risk Management

Model risk management (MRM) is an established practice with renewed relevance. It arose in the wake of the financial crisis when the Federal Reserve Board and other financial services regulators made it a required safeguard for models. But now, in the era of Generative and Agentic AI, MRM can bring the same kind of confidence to the policies, procedures, governance, and controls to help keep the latest AI applications trustworthy and reliable.

Every AI application rests on the foundation of a model designed to analyze data and make decisions. Some models are widespread and public; some are trained to focus on specific needs within organizations; some are closely held intellectual property (IP). But trust in any AI system starts with trust in the model, and testing and validating the model for use is a specialized discipline.

A critical safeguard with broader relevance

Validating an AI model for reliability, accuracy, and trustworthiness is a collection of processes, not a single one. It includes monitoring, reporting, and other quantitative and qualitative steps, along with organizational policies and procedures to guide and enforce their use.

Because of its origins in regulating risks such as market, liquidity, fraud, and credit risk, MRM is a more familiar discipline to financial services industry (FSI) organizations, even in its new AI centric form. Leaders in other industries may have more of a learning curve to embrace the way it works and why it's important.

In some industries, the need for it may come as a brand new challenge: Those such as energy, health care, life sciences, manufacturing, or consumer packaged goods have regulations of their own; but MRM in its current form would likely not have been among them. As with AI in general, model security is varied from sector to sector and organization to organization, with a wide span of experiences and maturity levels.

Stopping risk before it starts

A model that has been validated insufficiently, or not at all, can invite risks such as data and privacy breaches, hallucinations, bias, IP infringement, or vulnerability to attack. In turn, failures like those can lead to negative business outcomes such as financial and reputational loss, regulatory sanction, operational disruption, or strategic shortfalls. Within agentic AI structures, feedback loops and multi agent dependencies add to the risks.

Historically, assessing and managing model risk has been a safeguard that could be applied at the end of a development process. In the age of AI, the earlier the better to avoid having to reconfigure what might be months of work. No matter where an organization is along its AI journey, however, the least advantageous time to test a model is tomorrow.

Tools that bring Model Risk Management to life

Deloitte's involvement with MRM for AI models began with early investments that mirrored its overall commitment to AI leadership and extended its experience from earlier FSI needs. Its MRM team includes mathematicians and data scientists working closely with business and AI professionals to tackle model trust from all sides. Organizations have turned to Deloitte for MRM assistance to validate the operation of chatbots, search and summary tools, agents, and other AI uses.

The Deloitte approach to model validation and monitoring includes distinct steps to help mitigate risk:

- **Data assessment** of sources, quality, relevance, pre processing steps, and controls
- **Conceptual soundness** testing of a model's design, framework, architecture, and prompts, including the use of external tools and the configuration of AI agents as appropriate
- **Performance testing** to evaluate and quantify different risks through actual operation, including hallucination, accuracy, skills, and other factors
- **Implementation** of processes and controls including qualitative review, memory assessment, and user acceptance testing (UAT)
- **Ongoing monitoring** of selected metrics to track performance, user interaction, ethics, and other key indicators

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.

Clifford Goss, PhD
Partner
AI Leader, Financial Services
Deloitte & Touche LLP
cgoss@deloitte.com