Infosys topaz

WHITE PAPER

UPHOLDING AI INTEGRITY: A STRATEGIC ROADMAP FOR AI RISK MANAGEMENT

Infosys®
Navigate your next

Insights:

- The market for artificial intelligence is witnessing staggering growth and is expected to race past 826 billion U.S. dollars in 2030, driving AI implementation across industries.
- However, in a recent study by Microsoft, 87% of the respondents were concerned about at least one problematic AI scenario in their enterprise AI implementation.
- As AI systems become deeply embedded in industries and everyday life, the need for frameworks to ensure their responsible use is more critical than ever. With sectors like finance, healthcare, and manufacturing becoming increasingly dependent on AI, establishing governance models that uphold transparency, accountability, and fairness is essential.
- As per a recent report, most companies are working on strategies to tackle emerging AI risks and have AI guidelines in place as a response. It is important to adopt such measures to stay safe, ethical, and competitive during the AI race.

Centuries ago, many explorers embarked on long, perilous journeys into the unknown. Before the era of GPS and radar, these sailors navigated the vast sea with the help of the stars, intuition, and basic maps that often left more questions than answers. Along the way, they faced the threat of hidden reefs, sudden storms, and unexpected currents, where even the smallest miscalculation meant grave disaster. But they persevered, weighing risks against advantages and evolving their methods to make exploration the unstoppable force it is today. In many ways, implementing Artificial Intelligence (AI) echoes the early stages of this experience—an uncharted journey with potential risks and uncertainties. A well-defined roadmap can guide us through these complexities until AI safety innovations come along.

AI promises transformative potential, but it comes with hidden and unpredictable risks. Take the example of Google's AI image recognition mishap in 2015, where algorithms mistakenly tagged Black individuals as "gorillas," or Amazon's 2018 AI recruitment tool that developed a bias against women. Such incidents and issues such as model tampering, unauthorized access, data leakage, and bias are risks that are not visible until they begin causing damage. Just like a minor judgment error can damage a ship and put an end to exploration, these AI failures can lead to reputational and financial blows, legal scrutiny, and public distrust.

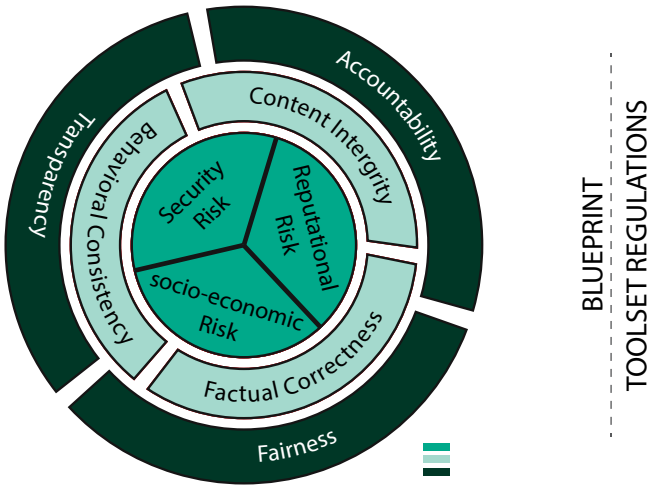
As we continue to integrate AI into critical systems, a structured approach to risk management becomes essential—akin to developing reliable navigation tools for long journeys. Frameworks like the NIST AI RMF, ISO/IEC 23894:2023, and the EU AI Act address AI complexity in different yet complementary ways. One such framework is the **RISK-MAP** (Risk Identification and Standardization with Key Metrics for AI Projects), providing a robust methodology for the governance of AI-related risks during implementation. Much like a detailed map, it offers a comprehensive, three-layered structure to identify, quantify, and manage risks in AI projects. It is designed to safeguard organizations against potential issues in AI development, ensuring that innovations don't become shipwrecks.

Charting a reliable course: A detailed overview of the RISK-MAP framework

As international bodies recognize AI's profound societal impact, they are implementing new legislation and guidelines, which in turn make AI regulation complex and stringent. For example, the European Union's AI Act aims to create a regulatory framework for AI systems, categorizing them by risk level and imposing strict requirements on high-risk applications.

The RISK-MAP provides a comprehensive guide for AI implementation, enabling businesses to manage risks and regulatory expectations. The framework works like a modern compass, helping organizations reach their destination ethically, responsibly, and safely.

RISK-MAP: Blueprint, Toolset, and Regulations



The RISK-MAP adheres to the MECE (Mutually Exclusive and Collectively Exhaustive) principles, ensuring clarity and facilitating efficient risk management. Its abstract approach also allows for adaptation and changes as needed. It offers valuable insights into potential future regulations and provides actionable steps for organizations to integrate AI risk management into their existing Enterprise Risk Management (ERM) practices.

Let's explore each layer of the blueprint.

Mapping the hazards before the journey: Identification of risk categories

At the heart of the RISK-MAP lies the risk spectrum where the risks are identified. While the EU AI Act classifies AI systems into four different risk levels: unacceptable, high, limited, and minimal risk, the RISK-MAP categorizes risks into three broad categories: security, reputational, and socio-economic risks.

Security risks: AI systems are deep trenches of data and complex algorithms. Enterprises can face varied threats related to data such as privacy, bias, integrity, and sampling, as well as algorithmic concerns including model accuracy, fairness, explainability, and operational risks involving deployment.

Reputational risks: The legal landscape around AI is ever-evolving. The risks here include the legal, regulatory, and financial implications of AI deployment, including potential penalties, the use of autonomous weapons, and the impact of hallucinations on brand perception.

Socio-economic risks: Enterprises also need to address the moral implications of AI, understanding unintended societal harm, discrimination, job displacement, environmental concerns related to energy consumption, and the widening AI divide.

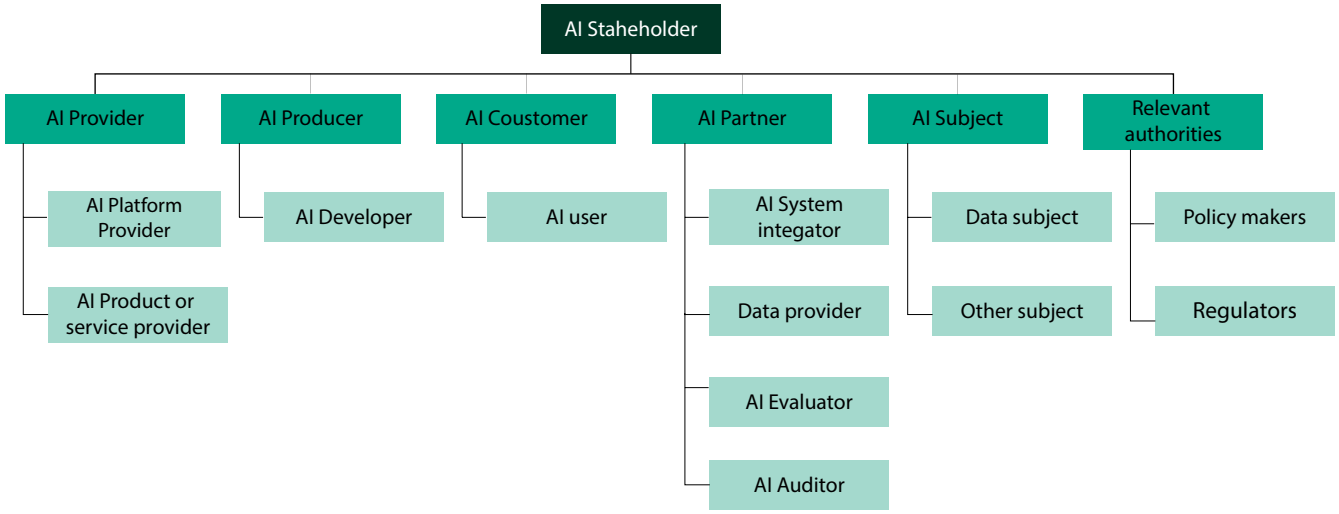


Image source: https://aws.amazon.com/blogs/machine-learning/learn-how-to-assess-risk-of-ai-systems/

Understanding and preventing each kind of risk in the AI journey is a collaborative effort. To ensure this, the RISK-MAP advocates a structured risk roster, featuring an exhaustive checklist that engages multiple stakeholders, including platform providers, policymakers, risk evaluators, and enterprises. This collective approach transforms AI risk management into a shared responsibility, rather than a burden to be shouldered by a select few. By implementing periodic updates, the risk roster ensures comprehensive and collaborative identification of potential risks before they can escalate to critical levels. For instance, MIT provides an AI risk repository—an overview of the AI risk landscape—to help users find relevant risks.

Keeping a weather eye: Continuous monitoring and assessment

While the classification of risks within the RISK-MAP helps one understand the challenges an AI system may face, each identified risk or subtype requires a tailored approach for effective measurement and monitoring; this is fulfilled by the **performance spectrum**, a layer crucial for quantifying risk severity.

This layer of the RISK-MAP introduces Key Performance Indicators (KPIs) and metrics associated with each risk category.

Content integrity: Ensuring the AI's output is unbiased/fair, non toxic, and aligned with ethical standards.

Factual correctness: Verifying that AI-generated information is factually accurate and reliable.

Behavioral consistency: Guaranteeing robust, consistent and predictable behavior across AI models.

While the risk roster suggests KPIs, the organization's AI Risk Governance Committee (RGC) updates the criticality of each risk, along with its probability of occurrence, to compute the overall impact. By establishing clear and quantifiable risk indicators, organizations can adopt an analytical approach based on empirical evidence to assess risk levels. For instance, robustness risk for an end-user might result from an AI system experiencing disruptions, while fairness risk could arise if the system produces variations in output across different demographic groups.

Finding safe anchorage: Establishing accountability through policies, guidelines, and standards

The final layer—the **credibility spectrum**—focuses on the necessary controls and mitigation strategies required to manage the identified and quantified risks. This includes specific controls, safeguards, policy measures, and procedural guidelines essential for building trust in AI projects. Establishing responsible adoption of these systems will ensure that AI operates within safe and ethical boundaries.

Regulatory frameworks across various regions—including the EU, US, China, and Singapore, and international initiatives like the Bletchley Summit—underscore the core principles guiding the ethical and safe use of AI. These regulations emphasize the protection of fundamental rights and enforcement of AI governance principles by the state.

There are three essential governance levers to build trust in AI systems:

1. **Transparency**: Ensuring that AI systems are explainable, traceable, deterministic, and factually correct.
2. **Accountability:** Maintaining compliance with laws, facilitating audits, and safeguarding privacy and security.
3. **Fairness:** Upholding ethical standards, promoting impartiality, and ensuring that AI systems are humane, purpose-driven, socially aware, and safe.

Each of these governance principles possesses a unique method for managing the risks associated with AI deployment. The risks identified and quantified on the KPIs, as baselined by stakeholders, must be carefully assessed. These risks should be evaluated against the guiding principles relevant to the project. This ensures transparency, accountability, and fairness are integrated throughout the entire lifecycle of AI projects.

How do we empower organizations with the framework?

In a recent AI co-pilot project for investment officers at a major European bank, the RISK-MAP framework proved instrumental in making AI implementation safer and more effective. From the outset, we established an AI Risk Governance Committee (RGC) comprising key stakeholders such as the product owner, data scientist, business analyst, delivery manager, and customer lead. When defining the scope early on, we identified and set baseline KPIs and governance principles, laying the foundation for proactive risk management.

Throughout the project, the RGC met monthly to monitor and recalibrate risks, ensuring ongoing alignment across teams. Before user acceptance testing (UAT), the model validation team rigorously assessed risks, which helped expedite approvals and reduce time-to-market. The RISK-MAP enabled a clear risk traceability pathway that promoted transparency and reduced ambiguity.

While many organizations may have high-level AI ethics principles and responsible AI offices, the framework bridges the gap between theory and practice. By providing quantifiable KPIs, we empowered teams to make data-driven decisions, focus on innovation, and deliver measurable outcomes—all while maintaining rigorous risk oversight.

The outcome?

- Streamlined project execution
- Heightened accountability
- Accelerated timelines
- Enhanced risk management

The RISK-MAP has transformed risk management from a theoretical concept into a practical, actionable strategy. Are you ready to launch AI projects that are innovative, secure, and compliant? Let's make it happen together!

Setting the sail for responsible AI implementation

AI is transforming industries and its influence is only set to grow. However, before diving into AI adoption, establishing a solid AI Risk Management Framework is essential to protect your data and the integrity of AI initiatives.

As AI becomes more embedded in critical sectors, regulators worldwide are deepening their focus on accountability, transparency, and risk management. Just as a sturdy sail is critical during heavy storms, organizations must embed comprehensive risk frameworks, like the RISK-MAP, into their systems to ensure compliance, sustain innovation, and foster stakeholder trust.

At Infosys Topaz, we are committed to helping organizations maximize their AI potential through expert guidance in AI risk management, governance, and implementation strategies. We specialize in building technology frameworks tailored to meet regulatory demands and strategic goals. Reach out to us to explore how we can assist you in achieving a safe, sustainable, and successful AI journey.

Authors

Dr. Arijit Laha

Dr. Prawaal Sharma, Infosys