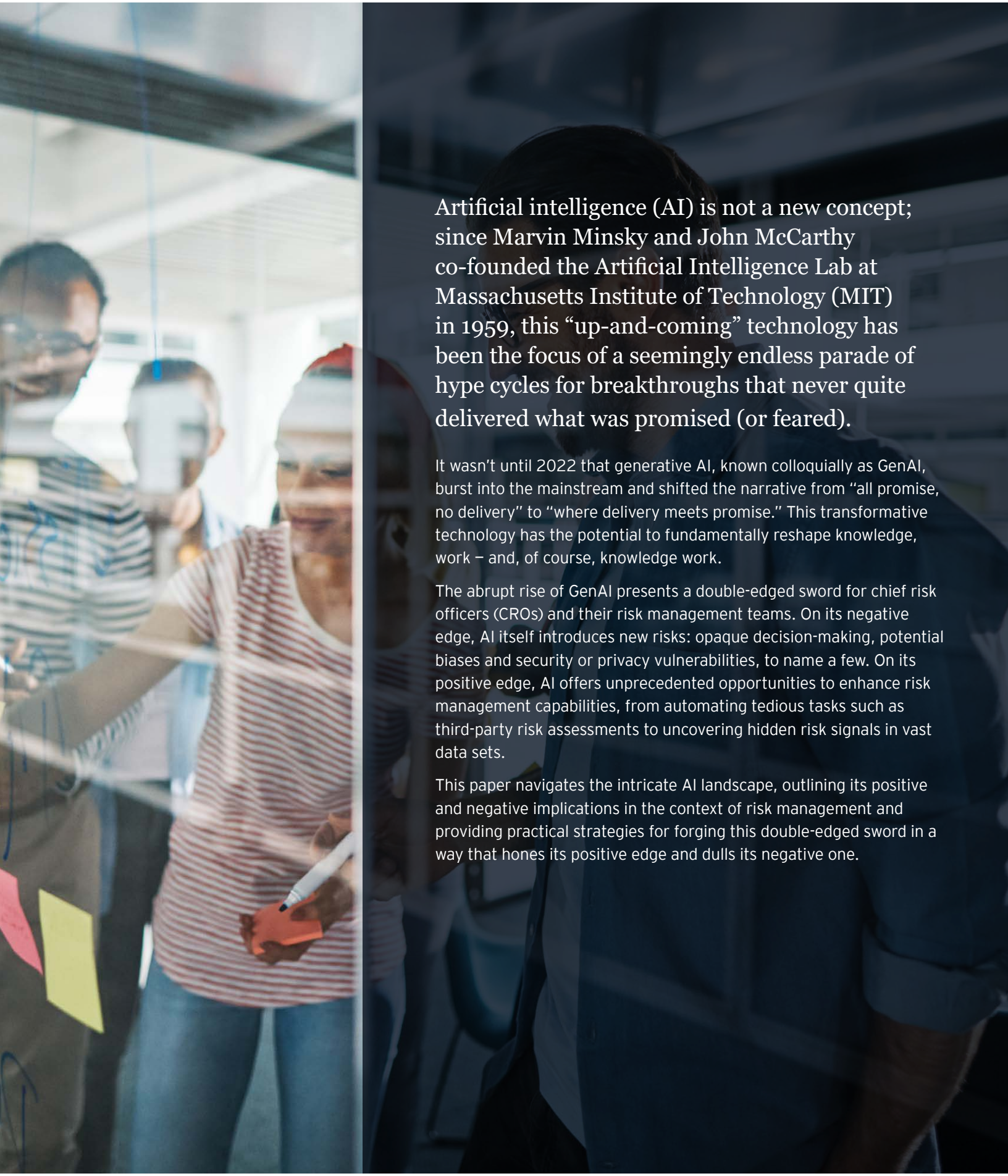
Wielding the double-edged sword

EY
Building a better working world

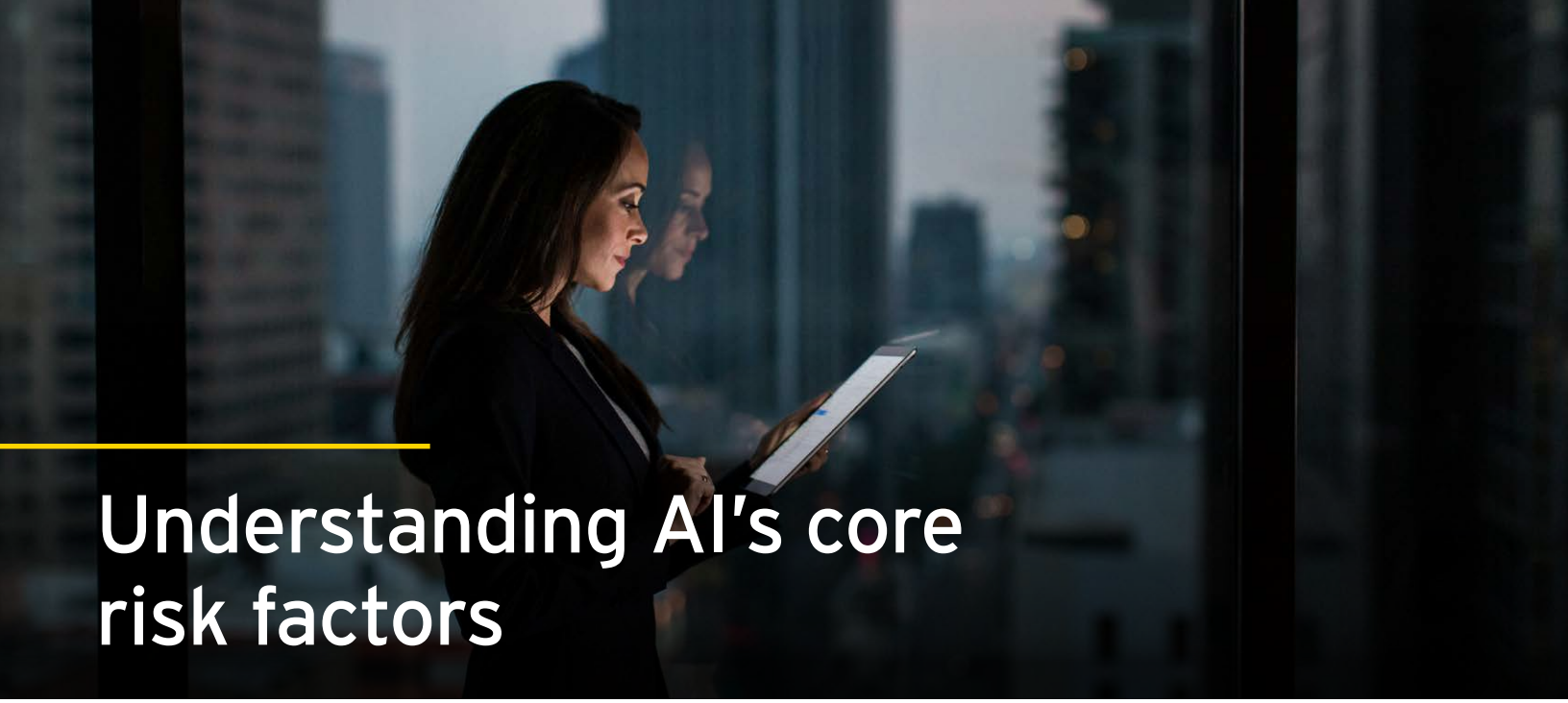


Artificial intelligence (AI) is not a new concept; since Marvin Minsky and John McCarthy co-founded the Artificial Intelligence Lab at Massachusetts Institute of Technology (MIT) in 1959, this "up-and-coming" technology has been the focus of a seemingly endless parade of hype cycles for breakthroughs that never quite delivered what was promised (or feared).

It wasn't until 2022 that generative AI, known colloquially as GenAI, burst into the mainstream and shifted the narrative from "all promise, no delivery" to "where delivery meets promise." This transformative technology has the potential to fundamentally reshape knowledge, work — and, of course, knowledge work.

The abrupt rise of GenAI presents a double-edged sword for chief risk officers (CROs) and their risk management teams. On its negative edge, AI itself introduces new risks: opaque decision-making, potential biases and security or privacy vulnerabilities, to name a few. On its positive edge, AI offers unprecedented opportunities to enhance risk management capabilities, from automating tedious tasks such as third-party risk assessments to uncovering hidden risk signals in vast data sets.

This paper navigates the intricate AI landscape, outlining its positive and negative implications in the context of risk management and providing practical strategies for forging this double-edged sword in a way that hones its positive edge and dulls its negative one.



Understanding AI's core risk factors

The risks of AI are numerous and demand careful consideration. To understand these risks, it's useful to consider the three core factors that underpin them: pace, innumerability and magnitude.

The pace factor

The phrase "gradually and then suddenly"[1] has been used to describe any number of technological and socioeconomic developments, from bitcoin to bankruptcy. But it's an especially fitting description of the ascendance of AI, which reached an inflection point in 2023. This extreme pace has its origins in several conditions: a confluence of technologies and developments surrounding AI, with the internet, mathematics, communications and social media acting as amplifiers; the intrinsic ability of AI to evolve in a self-referential way – essentially improving itself; and the widespread adoption of AI by developers to rapidly create new AI products. This pace shows no signs of slowing; in fact, it's likely that these conditions will act as a flywheel, further accelerating the pace of AI evolution to an unprecedented degree.

While thrilling for technologists and revelatory for businesses, the pace factor introduces a challenge for risk management teams. This can manifest in a variety of ways, such as:

- **Rapid obsolescence of risk management models.** The swift evolution of AI technologies can render existing risk models and mitigation strategies obsolete at an astonishing rate, leading to scenarios in which risk management teams are perpetually playing catch-up, unable to effectively anticipate or respond to new AI-related risks as they emerge.
- **Overreliance on automated systems.** As AI development accelerates, there's a tendency for organizations to increasingly rely on automated systems for decision-making. This overreliance could lead to a lack of human oversight, making it difficult to identify and correct errors or biases inherent in AI algorithms.

Ultimately, the rate of change will be unmanageable, and techniques to manage the resultant risk will need to be created as quickly as the technology itself develops.

The innumerability factor

Unlike most technologies, which have well-defined use cases that imply a well-defined scope, AI was not developed to address a particular use case; it was developed to reproduce humanlike intelligence at scale. The implication is that AI has an unbounded number of use cases, and therefore, an uncountable number of risks. This ambiguity is exacerbated by AI's lack of deterministic outcomes; reproducing a risk is infinitely more difficult when the context is unbounded. For example, organizations may face:

- **Unpredictable AI applications and consequences.** Given the broad scope of AI applications, it becomes challenging to predict how AI might be used (or misused) in various contexts. This unpredictability makes it difficult for risk management teams to foresee and prepare for all potential risks, especially those arising from novel or unintended uses of AI.
- **Difficulty in regulatory compliance.** The diverse applications of AI across different sectors can lead to complex regulatory challenges. Ensuring compliance with varying and evolving regulations across different domains becomes a significant tactical risk, given the legal and financial repercussions associated with noncompliance.

The innumerability factor means that risk management teams will need to manage risks categorically rather than using a "point solution" approach.

The magnitude factor

Historically, humans — whether intentionally as bad actors or unintentionally through error — have been the risk vectors that have led to the most drastic consequences. Because AI not only approximates a human's ability to reason but is also intended to be autonomous, it stands to reason that AI will at least match the magnitude of humans as a risk vector. In fact, given the innate scalability of AI, it is not unimaginable that it will supersede humans as a risk; after all, human risk is intrinsically bounded by a finite number of people, while AI is not. The magnitude factor can manifest in many ways, such as:

- **Enhanced scale of cybersecurity threats.** As AI systems become more capable, they also become more attractive targets for cyber attacks. The magnitude of potential damage from such attacks is amplified, considering AI systems may control critical processes and often have access to vast stores of sensitive data.
- **Escalation of operational disruptions.** The increasing autonomy and complexity of AI systems can lead to significant operational disruptions that eclipse those caused by traditional system failures. When a highly interconnected, mission-critical AI system malfunctions or is compromised, the impact can escalate from local to global very quickly, resulting in substantial financial losses, damage to reputation and erosion of stakeholder trust.

To counter the magnitude factor, risk management organizations must adopt a more holistic and proactive approach that incorporates advanced predictive models and continuous monitoring systems that can adapt and respond to the rapidly evolving AI landscape. This approach should focus not only on mitigating known risks but also on anticipating and preparing for emergent risks that AI's autonomous and self-improving capabilities could generate.

Managing AI risk

While the risks posed by AI are intimidating, they are not insurmountable. Building a holistic foundation according to the principles of "systems thinking,"[2] as popularized in Donella H. Meadows' *Thinking In Systems: A Primer*, is the key to safe and responsible deployment of AI at scale.

This foundation consists of a three-pronged AI risk management system that (a) establishes a "collective mental model" for thinking about AI risk, (b) leverages that mental model to create an AI risk management framework, and, most importantly, (c) taps into AI itself to scale risk mitigation at the levels described in the three core risk factors.

Establishing a collective mental model

Any successful risk management system depends on the risk team's mindset and ways of working, and an AI risk management system is no different. To make the best, most timely decisions, the risk team must comprehensively understand the dynamics of AI (including its capabilities, limitations and ethical considerations) and operate in an environment that encourages the continuous learning and adaptability that the fast-paced AI landscape demands. A shared mental model achieves both of these goals.

Typically, this is best established by creating five to 10 irreducible, immutable **written principles** that the team applies, by default, to resolve ambiguity or internal disagreement. This "principle-oriented alignment" ensures that the team can collaboratively and proactively address AI-related risks rather than reacting to them in a disjointed or piecemeal fashion.

Creating an AI risk management framework

Once a mental model is established, creating an action-oriented risk management framework becomes possible. This framework should be dynamic, scalable and able to evolve with the rapid developments in AI, using mechanisms for regular review and updates. It should encompass a comprehensive risk identification process, robust risk assessment methodologies and effective risk mitigation strategies.

Finally, the framework should integrate cross-functional collaboration, bringing together expertise from various departments such as IT, legal and operations to provide a holistic view of AI risks and their management.

Tapping into AI as a partner in managing risk

Consider this: Would a chief information officer (CIO) trust a team to manage cybersecurity without the same tools that hackers use? Likely not. Given the scale of risk, this principle is especially true in the AI era. To put it simply, managing AI risk without AI tools is akin to trying to extinguish a forest fire with a medicine dropper full of water. That's why leveraging AI to manage risk is arguably the most important prong of them all.

Unlike the first two prongs of an AI risk management system, this prong will be uncharted territory for risk teams and requires an innovative and forward-thinking approach that recognizes the ways AI can automate and enhance risk management.

By turning AI into a partner, organizations can not only keep pace with the rapid developments in AI technology but also harness its power to create more efficient and effective risk management processes.

For example:

- **Proxying access to core AI services.** Instead of granting unrestricted access to external AI tools, organizations can establish secure proxies that filter/sanitize input and output data. This "firewall for AI behavior" allows employees to leverage the benefits of AI while ensuring data security and compliance.
- **AI monitoring and alert systems.** AI decisions must be continuously monitored to identify anomalies and potential biases. Implementing AI-powered alert systems can notify human monitors when outputs deviate from acceptable thresholds, enabling timely intervention and course correction.
- **Data verification and cleansing systems.** Data used to train an AI model must be verified and cleansed. Automated tools can identify and remove biases and inaccuracies, while machine learning algorithms can be used to detect and rectify data anomalies in real time.
- **Self-regulating ethical AI frameworks.** Developing AI models that incorporate ethical guidelines and industry benchmarks can help mitigate the risk of biased or harmful decisions. These models can be designed to auto-detect potential ethical violations and trigger feedback loops for automatic correction or human intervention.
- **Leveraging a "virtual risk manager."** While it may seem like science fiction, AI technology is nearly mature enough to create fully autonomous AI agents with reasoning comparable to (and in specialized circumstances, better than) that of a human. The ultimate goal for any CRO should be a "virtual risk manager" that can automate risk assessments and remediations.

Conclusion

The inexorable rise of AI, and its inevitable integration into organizational frameworks, presents a double-edged sword of formidable challenges and remarkable opportunities. To ensure the technology is a force-multiplier and not a vector for unmanageable risk, it must be approached strategically.

The onus is on CROs and risk management teams to develop a nuanced understanding of AI's risks and benefits and to cultivate an environment that balances utility and innovation with vigilance. An AI risk management system built on a sophisticated framework, designed to evolve in tandem with the technology and to tap into the power of GenAI to act as a risk management partner, will enable teams to unlock its benefits: enhancing and streamlining processes, performing data analysis, predicting trends, supporting decisions, and more.

The roadmap for successfully integrating AI into risk management includes the establishment of robust AI governance frameworks, continuous monitoring and updating of AI systems, and fostering a culture that balances technological reliance with critical human oversight.

The recommendations for organizations are clear:

1. Develop and continually refine a comprehensive AI risk management framework, guided by a well-defined mental model.
2. Use AI not just as a tool but as a partner in risk management, employing its strengths for enhanced efficiency and effectiveness.
3. Implement proactive measures such as AI monitoring and ethical AI frameworks to mitigate risks associated with AI integration.

Millennia have taught us that a sword can be a weapon or a tool; how one forges (and wields) it makes all the difference. The double-edged sword of AI is no different. Embracing AI in risk management is not just about navigating its challenges but also about unlocking its potential to revolutionize how risks are identified, assessed and mitigated. As AI continues to evolve, so too must the strategies employed to hone it, ensuring that organizations not only keep pace with technological advancements but also harness them to maximize success and resilience in an increasingly complex world.

The views reflected in this article are the views of the author and do not necessarily reflect the views of Ernst & Young LLP or other members of the global EY organization.

References and citations

1 Hemingway, Ernest. The Sun Also Rises. 1926.

2 Meadows, Donella H. Thinking In Systems: A Primer. Chelsea Green Publishing, 2008.

Contacts

Sinclair Schuller
Managing Partner
Nuvalence

Samta Kapoor
Partner
Ernst & Young LLP

Kapish Vanvaria
Partner
Ernst & Young LLP

EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

US SCORE no.
2402-4422041
ED None

ey.com