Guidance for the commercial stakeholder

How to harness AI benefits and mitigate risk

Guidance for the commercial stakeholder: Harness AI benefits and mitigate risk

Anekanta®

Prepared by: Anekanta®AI and Anekanta® Consulting
Issued 10th August 2024 For external distribution

Anekanta®AI and Anekanta®Consulting

Contents

1. Introduction

The rapid advancement of AI has ignited both excitement and apprehension. While the potential benefits are immense, so too are the risks. Unlike other technologies, AI presents a unique set of challenges that demand tailored approaches.

In this short article, Anekanta® provides a range of insights into the AI risk considerations that need to accompany AI opportunity decisions, discussing challenges such as identifying AI systems in use and explaining the complexity of AI risk. We cover issues like the black box problem, data, interconnectedness, and the rapidly evolving AI risk landscape, concluding with remedies and actions organizations can take to mitigate risks originating from multiple covert sources.

This article is intended for busy commercial professionals responsible for AI governance and risk management, including IT leaders, board members, and procurement teams, offering insights and actionable steps to help organizations navigate the complex landscape of AI risk to achieve trust in AI.

2. AI is more than Gen AI

A common misconception is that AI is synonymous with generative AI. This is the technology behind tools like ChatGPT, Gemini and Claude. These tools are adding value in industry when implemented securely and when proprietary databases are used, they have improved accuracy. While undeniably powerful, generative AI is just one facet of a much broader spectrum. At the core of many established industrial systems are classical AI techniques. These systems range from aviation, manufacturing, critical national infrastructure to security. Techniques such as machine learning often operate invisibly and this ‘hidden AI’ can be equally, if not more, impactful when developed or used irresponsibly. For example, a government welfare system reliant on AI algorithms can produce biased outputs. These outputs cannot be challenged and can have catastrophic consequences, as seen in the Dutch Government child welfare scandal.

3. The complexity of AI risk

The complexity of AI risk stems from several key factors, here are some examples which are by no means exhaustive:

Black box problem

Unlike traditional software, AI models often operate as ‘black boxes’, making it difficult to understand the decision-making process. This opacity hinders risk assessment and mitigation. Additionally, AI algorithms have been designed in and buried within other systems. Their idiosyncrasies are revealed through other effects such as social impact. Examples include the automated and targeted delivery of misinformation and disinformation.

Data dependence

AI systems do not work without data. Biased or inaccurate data can lead to discriminatory or harmful outcomes therefore identifying and addressing data quality issues is crucial. Good data governance practices throughout the AI system lifecycle are essential. They will reduce the risk of biased, unreliable data from entering the AI training cycle or model. Standards such as ISO/IEC 27001 and ISO/IEC 27701 can reduce AI risk in the procurement process.

Guidance for the commercial stakeholder: Harness AI benefits and mitigate risk

A growing challenge where AI risk will increase is when a combination of expert systems are integrated with GenAI interfaces to improve human-machine interaction. Without due care and attention, anonymised, pseudonymised, and ring-fenced data when combined may re-identity individuals and enable biometric categorisation. The misuse of personally identifiable information (PII), sensitive, and biometric data increases the risk of regulatory penalties. It also leads to reputational damage to the organization. In the EU, penalties can be up to 7% of total worldwide turnover for the previous financial year or EUR 35 million under Article 99 of the EU AI Act.

Unintended consequences

AI models can exhibit unexpected behaviours, leading to unforeseen consequences. This is particularly challenging in safety-critical applications. These require accuracy derived from deterministic decisions rather than the probabilistic 'best guess' synonymous with GenAI. AI systems can produce unintended consequences when asked to perform tasks outside the realm of the training data. For example, a facial recognition system whose algorithm is trained on data from predominantly white male images will produce unreliable results when exposed to any other demographic.

Developers in this domain have been working on this problem to extend training data sets, often using synthetic data to improve its diversity. Additionally, users are starting to recognise that they must assess AI risk in the context of the organization's use. They also need to institute robust AI governance policies and practices. This will reduce the instances of bias amplification due to the way the software is used.

Interconnectedness

AI systems are increasingly interconnected, creating complex dependencies and potential cascading failures. This issue is not adequately considered due to an over-focus on human productivity gains through the use of GenAI and chatbots. There are far more autonomous interconnected devices (IoT) on the planet than people. Estimates range from 25 to 75 billion, with a midpoint identified in research, of nearly 40 billion by 2033. The interconnectedness of critical sectors can lead to new AI risk scenarios. These scenarios model the potential that IoT devices may be recruited into a AI generated, self-propagating attack.

A rogue application recently disrupted critical infrastructure, experienced in the Crowdstrike cybersecurity software update. This update caused Microsoft systems to shut down because of an unexpected error it was unable to process. This issue revealed the vulnerability of a range of operating systems whose cores are usually well protected from external access. However, they could be exploited by a string of GenAI agents working together. This is especially true for agents that can effectively mimic trusted humans, such as deepfakes, which allow external access through deception.

Rapid evolution

The AI landscape is evolving at an unprecedented pace, making it difficult to keep up with emerging risks. Models which have been designed into the value chain may change. This change results in a range of emerging AI risks. These risks need to be understood, mitigated and managed. Boards, IT and procurement leaders need to know what they are purchasing. They need to understand not just the range of opportunities but also the consequences which may arise. The AI risk landscape is more complex than any other technology innovation. This is due to AI's ability to learn from data and make decisions autonomously.

The additional risk of anthropomorphisation may lead humans to put too much trust in the machine. Additionally, the propensity for humans to defer to the machine is another factor. This is seen in the use of facial recognition technology and also manifested in the UK's Post Office scandal. These examples step into the realm of human factors extending beyond the concept that AI risk is just about software. It is crucial that organizations assess AI risk as a continual process. They should not wait for an adverse event to occur before taking action.

To mitigate AI risk, organisations need to establish a different way of doing things. This new way is more hypothetical and cross functional in its approach to imagining potential risk scenarios. It is combined with tried and tested risk assessment techniques based on probability and severity, however, a deep understanding of both AI techniques and real-world applications is crucial.

4. The importance of understanding AI risk

Recognising the multifaceted nature of AI risk is essential for effective management. It is particularly important in the context of emerging regulations like the EU AI Act. The AI Management System Standard ISO/IEC 42001 is a good start point for any organization moving towards compliance. It helps in maximising AI opportunities and reducing AI risk through AI policy setting. The standard provides a structured approach to AI governance. However, its effectiveness depends on a deep understanding of the underlying complexities of AI. AI systems could use an infinite combination of AI techniques and may be utilised for different purposes across the business landscape. They range from facilities management to HR and customer service. They also include operational and safety systems in critical infrastructure, aviation, and transportation.

A one-size-fits-all approach to mitigate AI risk is not only ineffective but can also be counterproductive. It could stifle innovation or lead to excessive bureaucracy. Instead, a nuanced perspective is required, one that considers the specific context, the type of AI involved, and the potential impact of unintended consequences.

By acknowledging the intricate nature of AI risk and adopting a tailored approach, organizations can better protect themselves, their customers, and society as a whole. This requires a combination of technical expertise, ethical considerations, and robust governance frameworks. Ultimately, it is about striking a balance between harnessing the benefits of AI and mitigating its risks.

5. Where to get help

An independent, specialised responsible AI services company can be a critical asset in navigating the complex landscape of AI risk. By providing unbiased assessments and recommendations, these firms can help organizations identify, prioritise, and mitigate risks effectively. Their expertise in evaluating a wide range of AI technologies and applications allows them to develop tailored strategies. These strategies align with an organization's specific needs and objectives. Furthermore, they can help ensure that AI risk management initiatives are integrated seamlessly across different business functions. This fosters a cohesive and effective approach to AI governance.

6. The role of specialised Responsible AI service providers

In addition to generic risk guidance for example ISO/IEC 23894, EU AI Act and NIST AI RMF. Anekanta® has developed a range of specialised frameworks designed for specific technologies and industrial use cases. These frameworks solve the AI risk assessment problem for sectors and domains using high-risk AI systems. Examples include the AI Risk Intelligence System™ for high-risk AI and biometrics. This system has been featured in the UK Government Portfolio of AI Assurance Techniques and the OECD.AI Catalogue of tools and metrics for trustworthy AI. It is designed to assess and mitigate AI risk under the EU AI Act Chapter III and Annex III requirements for high-risk AI systems and biometrics.

Copyright and Intellectual Property Notice

Contact

Please direct feedback or enquires to AI-risk@anekanta.co.uk or visit the web site at www.anekanta.co.uk