The Need for Risk Management in AI Systems

holisticai.com/blog/need-for-risk-management-in-ai

AI Risk Management

February 27, 2023

Organisations are investing in AI tools and strategies to optimise their processes to gain a competitive edge. In some cases, entire industries are transitioning to a reliance on AI, which comes with a heightened risk that must be managed. As artificial intelligence (AI) technology continues to advance rapidly, organisations must be prepared to respond to the significant changes it will bring to their operations on a global scale. Indeed, it is essential to understand the risks associated with AI and the best practices for managing them to ensure a successful implementation of AI technology. This blog provides an overview of AI risk management, including a summary of the risks and best management practices.

What are AI risks?

On 26 January 2023, the National Institute of Standards and Technology (NIST), a leading voice in AI standards, released AI RMF 1.0, the Artificial Intelligence Risk Management Framework. AI RMF defines risk as "the composite measure of an event's probability of

occurring and the magnitude of its consequences". According to NIST, AI risks are the potential harms to people, organisations or systems resulting from developing and deploying AI systems. Examples of harm range from sexist hiring tools to uncontrollable trading algorithms causing market crashes. These risks can stem from the data used to train and test the AI system, the system itself (i.e., the algorithmic model), how it is used and its interaction with people.

Given AI systems' potential dangers, proactively monitoring AI-based products and services is essential. One way to achieve control and help ensure safety and security is by adopting a risk management solution to triage, verify, and mitigate AI risks.

Measure, Manage and Mitigate Risk

An effective AI governance, risk, and compliance process enables organisations to identify and manage risks. At a high level, AI governance can be broken down into three major approaches:

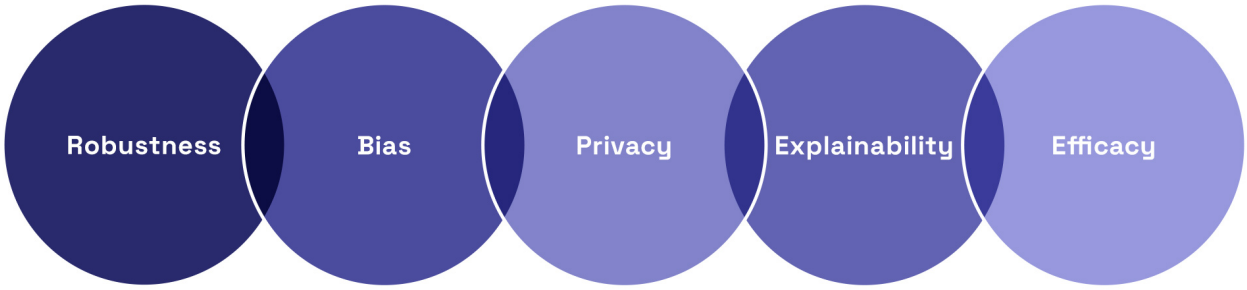
- **Principles** – using guidelines that inform and direct the use and development of AI, such as legislative standards and norms.
- **Processes** – to address risk and harm resulting from design issues and lack of appropriate governance.
- **Ethical Consciousness** – actions motivated by a moral awareness or desire to do the right thing. It encompasses the integration of codes of conduct and compliance, consideration of reputational issues, (corporate) social responsibility, and concerns for institutional philosophy and culture.

Ensuring that harmful or unintended consequences are minimised or do not occur during the lifespan of AI projects requires a comprehensive understanding of the role of responsible principles during the design, implementation and maintenance of AI applications.

What is AI risk management?

AI Risk Management is the process of identifying, assessing, and managing risks associated with using AI technologies. This includes addressing both technical risks (such as security vulnerabilities and algorithmic bias) and non-technical risks (such as ethical considerations and regulatory compliance). It involves understanding the potential risks and benefits of AI, developing strategies and policies to mitigate potential risks, and monitoring and responding to changes in the AI environment. Additionally, AI Risk Management also includes creating processes and systems to ensure compliance with ethical and legal standards, as well as internal and external policies.

When assessing a system, it is important to consider five main risk verticals:



1. **Robustness** is the risk of an algorithm failing in unexpected circumstances or under attack. It is essential to address when failure could result in financial losses or harm human well-being. This can be measured by assessing performance on unseen data and testing the system's ability to deal with targeted or adversarial attacks. Mitigation strategies for robustness risks include:

- Improving model generalisation.
- Retraining the model on new data.
- Using adversarial training and continual monitoring.

2. **Bias** risk is the risk that an algorithm mistreats individuals or groups and is particularly important for applications that significantly impact people's lives. Measuring bias involves looking at performance across different groups based on characteristics such as gender, ethnicity, and age. Data debiasing, model amendment, and output amendment can be used to reduce bias, depending on the source of the bias.

3. **Privacy** risk refers to the potential for an algorithm to leak sensitive or personal data. It is an important consideration for applications that process personal and sensitive data, as it can lead to data breaches and unlawful processing. Assessing privacy risk involves looking at the data type, the amount of data stored, and whether data minimisation techniques were applied. They can be addressed by reducing the training data, anonymising/pseudonymising data, or using de-centralised/federated models.

4. **Explainability** is the risk that the system or its decisions may need to be more understandable to users and developers. It is a key risk to consider when developing critical applications affecting many users. To reduce this risk, it is essential to examine the documentation and communication processes concerning models and data and how easy it is to interpret the model's decisions. Better documentation procedures can be developed, and tools can be used to interpret better the model's decisions, including how different features are weighted.

5. **Efficacy** is the risk that the system does not perform well relatively to its business case. It is a key risk to consider when working on projects where failure would have major consequences, such as a large financial loss. To reduce efficacy risks, it is important to measure the performance of the system using metrics such as accuracy, precision, and recall. Steps to improve model efficacy improving model generalisation, regularly monitor performance, and collecting additional training and test data.

AI regulation is coming, transparency is a first step

AI risk management will soon be codified by regulations that are being proposed around the world.

The European Union's proposed AI Act aims to create an 'ecosystem of trust' that manages AI risk and prioritises human rights in developing and deploying AI by adopting a risk-based framework to govern its use. In the United States, the White House has published a Blueprint for an AI Bill of Rights, which outlines the US government's vision for AI governance to prevent harm. To add, China has proposed a suite of legislation to regulate different applications of AI.

Providing a global explanation for an algorithm may seem straightforward. But organisations must make substantial structural changes in anticipation of AI implementation to ensure that their automated systems operate within legal, internal, and ethical boundaries.

Therefore, organisations with robust governance and risk management are best placed to ensure compliance with the increasing number of AI or use case-specific rules. Furthermore, by embedding a risk management framework, an organisation can move away from a costly, reactive, ad hoc approach to regulation.

Risk Management: the desire of many, the reality of few

Although AI adoption is soaring, risk management is lagging. The trouble is that many companies need help seeing that they have a problem. According to a report released by MIT Sloan Management Review and Boston Consulting Group, AI was a top strategic priority for 42% of the report's respondents. Still, only 19% said their organisation had implemented a responsible-AI program. The gap increases the possibility of failure and exposes companies to regulatory, financial, and reputational risks. While AI risk management can be started at any point in the project development, implementing a risk management framework sooner than later can help enterprises increase trust and scale with confidence. Advantages of AI Risk Management include:

- **Understanding of the AI inventory** – cataloguing AI systems provides insight into the systems being used in the company, how much has been invested into AI, whether there are any redundant systems and the types of AI being under-utilised and over-relied upon.

- **Upskilling the workforce** – in-depth assessments of systems and the resulting mitigation procedures result in a depth of knowledge about how the system works and associated risks. This will also necessitate new roles and responsibilities for those involved in the AI lifecycle, accountability frameworks, and AI risk training.
- **Improving the performance of AI systems** – testing an AI system's robustness and efficacy measures the system's overall performance level, meaning that steps can be taken to optimize underperforming algorithms.

Next steps

AI risk management will define the next era of technological advancement and become essential to companies' AI strategies. Given AI's rapid development and increasing applications, AI risks are constantly changing and evolving, meaning that comprehensive risk management strategies are needed to avoid reputational damage and facilitate legal compliance.

Auditing and testing AI systems reveal whether issues in a system's development, training, or deployment will lead to biased decision-making. Where any issues are found, these can be addressed with state-of-the-art mitigation techniques. In doing so, organisations can maximise their ability to innovate with confidence.

Interested in learning how to implement AI Risk Management strategy? Reach out to us at we@holisticai.com

Authored by Ayesha Gulley, Public Policy Associate at Holistic AI

DISCLAIMER: *This blog article is for informational purposes only. This blog article is not intended to, and does not, provide legal advice or a legal opinion. It is not a do-it-yourself guide to resolving legal issues or handling litigation. This blog article is not a substitute for experienced legal counsel and does not provide legal advice regarding any situation or employer.*

Share: