WHAT SHOULD WE KNOW ABOUT AI RISKS?

AI carries a range of serious risks alongside its many benefits.

Data-management and cybersecurity rules must establish a clear policy for the use of AI systems.

Organisations that share information with AI tools must train their staff so that they are able to recognise and mitigate AI-related risks.

Artificial intelligence (AI) offers significant opportunities for the public sector, private companies and individuals to create added value, improve workflows and foster innovation. However, these opportunities come with substantial risks that, if mismanaged or overlooked, can cause considerable harm to individuals, institutions and society as a whole.

AI-related risks require organisations that process information using AI systems, to think systematically, implement strong risk-management practices, and adopt clear policies and security measures. Mitigating these risks must occur at both the state and organisational levels by applying an AI-use strategy and organisational, ethical, legal and technical measures derived from it.

The primary risks associated with AI usage can be categorised into five types: data leaks, faulty training, cyberattacks, misinformation and manipulation.

The use of AI to handle high-risk, sensitive and especially classified information must be based on deliberate, well-considered decisions.

Data leaks pose a direct threat to the security of classified or other sensitive information, such as trade secrets. AI requires large amounts of data to train language models, and these may include sensitive, personally identifiable or strategically important information. Even when legal or contractual obligations prohibit disclosing such data directly, an AI system can still utilise information entered into it and, if prompted skilfully, may reveal it to third parties. In 2023, for instance, the tech world was shaken when Samsung employees entered trade secrets into a chatbot, potentially making the information accessible to unintended recipients. Another incident involved patient data left unsecured on a cloud server, allowing unauthorised access.

Preventing such leaks requires adherence to comprehensive information security management practices and the use of data protection and privacy measures such as pseudonymisation, anonymisation and data minimisation.

Pseudonymisation

Real individuals' data is replaced with aliases or codes that mask their identities. The original identity can be restored if the mapping key exists.

Anonymisation

Personal data is modified or deleted so it can no longer be linked to an identifiable person. This action is irreversible because no key or method exists to recover the original identity.

Data minimisation

The processing of personal data is limited to what is genuinely necessary for the intended purpose.

To prevent and detect data leaks, organisations must establish an AI-use policy as part of their information-management and cybersecurity policies. This includes setting access restrictions for AI systems and implementing logging, monitoring, and anomalies detection. It is advisable to follow core data protection principles even before deploying AI tools.

Faulty training is one of the most significant risks in developing large language models. When a model is trained on biased, inaccurate or sensitive data, its outputs may reproduce those issues. These errors are not merely technical; they can directly impact people's rights and undermine trust in the system.

Language models must be trained and deployed ethically and under supervision to prevent the leakage of classified or other sensitive information and to avoid harm to society.

For instance, Microsoft's experimental Twitter chatbot, Tay, began producing racist and sexist responses within just 24 hours of interacting with malicious inputs from users. As a result, the company had to take it offline almost immediately. Additionally, AI systems may produce fabricated answers rather than factual information because they are often designed to provide positive responses rather than acknowledge a lack of necessary information.

To reduce the risks associated with faulty training, it is essential to use high-quality, diverse and representative datasets. Data must be checked for quality and provenance, and the system's components must be examined for potential bias during analysis. Before deployment, models should be validated and tested across a range of scenarios. High-risk systems require continuous human oversight.

Cyberattacks pose a rapidly increasing threat as AI capabilities continue to advance. AI can be exploited for malicious purposes, such as identifying security vulnerabilities and creating malware. There are documented cases of AI-generated code evading traditional antivirus tools. Numerous reports note that AI is now employed at every stage of a cyberattack, from planning to execution.

To reduce the risk of hacking, organisations should implement technical security measures such as firewalls, encryption and detection systems. Regular security audits, vulnerability testing and attack vector analysis are also effective. As with any system, the timely application of security patches is crucial. Organisations should apply general cybersecurity risk-management methods and maintain continuous monitoring.

Misinformation is one of the most complex and dangerous aspects of AI for society. AI models can generate convincing text, audio and video that may contain factual errors or be deliberately misleading. In extreme cases, they can impersonate influential figures and present them as issuing instructions that are partially or entirely false.

The risk of misinformation is especially high during interstate conflicts, elections or crises. For example, AI-generated fake news has influenced election campaigns in Moldova and Ireland by creating panic over fabricated events. There are also cases where AI-generated content has pushed individuals to make irrational decision such as transferring money to criminals.

To reduce the risk of misinformation, organisations using AI systems must implement fact-checking mechanisms, content controls and deepfake-detection tools. It is also essential to identify and promptly remove accounts that spread false information, such as on social media platforms. Labelling AI-generated content and applying human oversight to assess the output accuracy are good practices. Finally, raising public awareness and strengthening critical thinking are vital to help people distinguish genuine information from deceptive content created with hidden intentions.

Manipulation refers to AI's ability to influence people's decisions, emotions and behaviour. The 2018 Cambridge Analytica case demonstrated to the public how algorithms can be used to build psychological profiles of voters and target them with tailored influence. AI can generate personalised content that may be emotionally charged and biased.

To mitigate such risks, transparency is paramount: every user should have a clearly defined right, and the practical means, to understand whether material, interactions or decisions directed at them are generated by AI and on what basis. Organisations using AI must raise awareness and, where necessary, provide training to help users recognise manipulative practices.

THREAT	COUNTERMEASURES
Faulty training	• Use validated and diverse datasets • Apply data filtering and pre-assessment • Avoid sensitive or biased content
Data leaks	• Remove classified information from any training data • Use sandbox environments and restricted access • Log and monitor model outputs
Cyberattacks	• Restrict the model's access to sensitive cyber infrastructure • Use application-level authentication • Conduct regular security audits and tests
Misinformation	• Implement fact-checking mechanisms • Use content filters and moderation • Train users to approach AI outputs critically
Manipulation	• Label AI-generated content for transparency • Limit personalised content recommendations in sensitive areas • Apply ethical content-creation guidelines

As AI solutions are adopted more widely, implementing cybersecurity measures at every level becomes critical. Using AI requires awareness, responsibility and carefully designed security measures. Only transparency and robust security can ensure that AI serves society rather than puts it at risk.