# Understanding the risks of AI and the frameworks guiding its safe use



In recent developments surrounding artificial intelligence, significant concerns have been raised about the risks inherent in AI models, particularly those related to data integrity, model accuracy, security, ethical considerations, and legal compliance. A notable incident in 2023 involved an Australian mayor preparing legal action against OpenAI after the company's chatbot, ChatGPT, generated a fabricated story falsely implicating him in a bribery scandal. This case highlighted the potential legal and reputational damages that AI-generated misinformation can cause.

Industry analysts, including Gartner, have predicted that about 85% of AI models will fail, principally due to issues with the quality and management of data. AI systems are fundamentally reliant on machine learning (ML) techniques, which require extensive training data to identify patterns and generate responses. Faulty or incomplete data can lead to errors such as wrong answers, biased outputs, and discriminatory decisions, underscoring the critical need for rigorous AI risk management.

AI risk is conceptually defined as the probability that an AI system will be exploited or produce erroneous results, multiplied by the impact of such errors. This risk varies according to context — minor errors, such as inaccurate historical dates for academic research, contrast starkly with high-stakes contexts like medical decision-making or autonomous vehicle operation, where inaccuracies could have serious or even fatal consequences.

AI risks are categorised into four primary types:

1. **Data Risks:** Challenges related to security breaches, privacy violations, data contamination, and biased data interpretation. A notable example is a healthcare risk prediction algorithm that showed racial bias by underestimating the care needs of Black patients due to flawed proxies based on medical spending patterns. This misinterpretation arose from systemic issues such as varying access to healthcare affecting spending, rather than patient health status.

2. **Model Risks:** Vulnerabilities inherent to the AI model’s logic or architecture, which can be exploited through adversarial attacks designed to confuse AI decisions, prompt injection attacks that manipulate generative AI outputs, or supply chain attacks targeting third-party components. Amazon's AI recruitment tool incident exemplified model risk where a male-dominated historic recruitment dataset led to sexist bias against female applicants.

3. **Operational Risks:** Problems arising from internal organisational factors including data drift (use of outdated data), inadequate scaling plans, improper system integration, and lack of accountability. The Apple Card controversy in 2019 serves as a key example, where algorithmic disparities in credit limits between men and women were compounded by a lack of transparency and human oversight, exacerbating the operational risk.

4. **Ethical and Legal Risks:** Potential violations of data privacy regulations such as GDPR and the California Consumer Privacy Act (CCPA), and failure to prevent discriminatory outcomes risk legal penalties and reputational harm. A high-profile case in March 2025 involved OpenAI facing a privacy complaint in Europe under GDPR after ChatGPT generated a false and harmful story about a Norwegian man, illustrating the growing legal scrutiny AI systems face.

In response to these multifaceted risks, several comprehensive AI risk management frameworks and standards have been developed:

* The NIST AI Risk Management Framework (AI RMF): Released initially in January 2023 by the National Institute of Standards and Technology, this voluntary framework emphasises trustworthy AI characteristics such as accuracy, safety, transparency, fairness, and data privacy. Its 2024 update, NIST-AI-600-1, specifically addresses generative AI risks, including misinformation and data privacy concerns such as unauthorised data scraping.
* The EU AI Act: Enacted in 2024, this is the first comprehensive AI regulation globally. It categorises AI applications by risk and imposes strict requirements, particularly for high-risk sectors like healthcare, banking, and recruitment. The Act mandates transparency on data use, prohibits unacceptable applications such as real-time biometric surveillance and social scoring, and empowers users to opt out of AI-driven decision-making.
* ISO/IEC Standards: These international standards, including ISO/IEC 27001:2022 and ISO/IEC 23894:2023, provide guidelines on data privacy, security, and fairness during AI development and deployment. An amendment in 2024 further incorporates privacy and sustainability into AI governance. The standards stress structured risk management, encryption, bias mitigation, regulatory compliance, and continuous security monitoring.

Given the complexities of data privacy and AI risk management, tools that automate privacy governance are increasingly important. Companies like Osano specialise in supporting organisations to comply with privacy laws, manage data ethically, and address AI-specific compliance challenges. Such platforms enable data sourcing, rights request processing, and risk assessment automation, contributing to safer AI deployments.

In summary, as AI technologies become more embedded in various sectors, managing the accompanying risks—ranging from data integrity to ethical use and legal compliance—remains paramount. Frameworks and regulations are evolving rapidly to address these issues, promoting accountability and transparency in AI systems while aiming to protect individuals and organisations from unintended harms.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.mossadams.com/articles/2025/02/data-security-and-integrity-for-ai-companies> - Discusses robust data integrity and security practices for AI systems, emphasizing compliance with regulatory standards and audit requirements.
2. <https://perception-point.io/guides/ai-security/ai-security-risks-frameworks-and-best-practices/> - Outlines AI security risks such as data breaches, adversarial attacks, and compliance measures, corroborating concerns about data integrity and model vulnerabilities.
3. <https://www.gooddata.com/blog/ai-security-in-data-analytics/> - Explains how AI security involves safeguarding ecosystems, ensuring model accuracy and regulatory compliance, aligning with discussions on data and operational risks.
4. <https://cloudsecurityalliance.org/blog/2024/03/19/ai-safety-vs-ai-security-navigating-the-commonality-and-differences> - Details integrity risks in AI ecosystems, including data/model poisoning and output manipulation, supporting claims about data and model risks.
5. [https://cybersecuritycrc.org.au/sites/default/files/2023-11/Poison%20the%20well%20-%20AI,%20data%20integrity%20and%20emerging%20cyber%20threats.pdf](https://cybersecuritycrc.org.au/sites/default/files/2023-11/Poison%20the%20well%20-%20AI%2C%20data%20integrity%20and%20emerging%20cyber%20threats.pdf) - Highlights data poisoning threats and societal impacts of compromised AI datasets, corroborating concerns about data integrity and legal risks.
6. <https://www.noahwire.com> - The original source, referenced for context on the article's claims, though specific content on the site is not independently verified here.
7. <https://news.google.com/rss/articles/CBMiigFBVV95cUxQYlJEMHF6NExZSWZSN2FpM2JxcnVPalpGQ0J4bXZkNUI5dmFlNWNBaThkZzdMdWFId0F4RmNEZGI2alBsQlRtdzlZTVZBN2o1M2h5UER4WmpVb2xhX2w0SUVCdlNzSEkzenRPZ3lSVWpFbFNXck9EUUZGZ0hpOXF1aVY5alZFdDJkeGc?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data