letsbloom

WHITE PAPER

AI Risk Management

An Approach to Manage Emergent AI Risks

Executive Summary

Rapid progress in the development and adoption of AI technologies in the past few years has brought about significant advancements across industries. However, these advancements have introduced new challenges and risks in how these technologies are used in business decision-making processes and what they mean from regulatory compliance and risk management perspectives.

This paper explores the key emergent risk and also outlines an approach to managing these risks enabling organisations to comply with existing and upcoming regulatory obligations.

The use of AI decision systems introduces risks around governance, transparency, explai nobility, robustness, i nclusivity, and interpretability. The regulatory environment pertaining to AI governance is steadily evolving, evidenced by the introduction of the EU's "The AI Act" and a corresponding Executive Order in the United States.

Organizations must consider and implement controls to comply with the regulations and manage these risks. This requires deliberate consideration during the adoption of AI technologies as retrofitting the controls after deployment may not be possible given the nature of the risks and the technologies involved.

Organizations should extend their existing Enterprise Risk Management Frameworks to include AI risk categories. This paper explores the key emergent risk and also outlines an approach to managing these risks enabling organisations to comply with existing and upcoming regulatory obligations.

Introduction

In 2012, Google's AI model demonstrated reliable identification of cats in images. However, progress in AI development remained gradual for several years. By 2019, the GPT-2 model, trained on 1.5 billion parameters and 8 million web pages, exhibited the capability to predict the next word in 40GB of internet text.

Fast forward to the present, GPT-4 has emerged, equipped to analyze any document and generate summaries based on contextual prompts provided by the user. Concurrently, other AI models can produce compelling image and video outputs, capable of deceiving most observers, reflecting the exponential pace of advancement in the field.

Whilst current models may lack some important capabilities that make them truly autonomous, as companies with large resources invest in the race towards creating Generative AI, a breakthrough seems inevitable.

At the same time, companies have started rapidly deploying these systems to support decision-making in ever-increasing critical business functions. The risks posed by these AI systems are amplified and new risks may emerge if organizations do not deliberately design their governance and risk management processes to handle these emergent risks.

Regulatory Landscape

In November and December of 2023, two significant AI regulations were introduced: '**The AI Act**' by the European Union and an Executive Order from the President of the United States of America. These documents offer valuable insights into the regulatory approaches taken by policymakers in the field of AI.

The EU's act focuses on the use of AI and specifically calls out

areas where the use of AI is prohibited (Title II of The AI Act, Prohibited Artificial Intelligence Practices). The us Executive Order outlines eight guiding principles and priorities (Section 2) that

focus on responsible development and use of AI technologies.

These regulations establish the foundation for compliance requirements for organizations. As such, organizations must align their risk management framework and governance processes accordingly.

The EU's AI Act mandates that an organization's risk management system should comprise a continuous iterative process conducted throughout the entire lifecycle of a high-risk AI system. entire lifecycle of a high-risk AI system.

This requires regular systematic updating for

I. Identification and analysis of known and foreseeable risks.

II. Estimation and evaluation of risks that may emerge when the AI system is in use and under conditions of reasonably foreseeable misuse.

III. Evaluation of other possible risks based on analysis post deployment of the system.

Similarly, the us Executive Order requires organizations to "Establish appropriate guidelines including appropriate procedures and processes, to enable developers of AI, to conduct AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems."

In addition, the order references NIST AI 100-l, which defines an AI Risk Management Framework to be implemented by organizations.

Key Emergent AI Risks

Regulatory requirements and prudent risk management practices dictate that organizations must identify and manage known, foreseeable, and emerging AI risks. Let us enumerate some of theserisks:

- **Governance:** Insufficient governance poses a risk by diminishing oversight across the development, deployment, and utilization of AI systems. As these systems advance in capabilities, there is a potential for exploiting the lack of oversight to introduce systemic biases or generate inaccurate outputs that elude controls.

- **Transparency:** AI models, due to their opaqueness in decision-making processes, often require extensive trial and error for verification. Regulators mandate specific transparency obligations for AI systems that engage with humans or have the capacity to impersonate or deceive humans through misrepresentation.

- **Explainability:** Even with adequate oversight and transparency, the underlying logic employed by AI systems, particularly deep learning models, can remain a black box and resist complete description. Thus, establishing robust guardrails during the learning process is imperative to manage risks effectively.

- **Robustness:** While AI systems may exhibit predictability within their operational parameters, their behavior can falter when confronted with novel situations or unpredictable inputs. Documented instances exist where advanced AI systems produce potentially hazardous responses to carefully devised malicious user prompts.

- **Inclusivity:** Biases in AI systems often arise from the datasets used for training and human biases/values introduced during the training process (e.g., reinforcement learning models). Regulators advocate for fairness principles to mitigate this risk, emphasizing the need for AI models to be inclusive, particularly with datasets that represent the entire population likely to use the system.

- **Interpretability:** Even a robust, inclusive, transparent, and explainable system might produce outcomes that can defy or confound expectations. Humans may need time to build trust in AI systems' logic, but that would be a slow process or may never fully occur.

Therefore, AI systems might still need checks and balances that can help interpret or override their decisions to minimize disruption.

Risk Management Approach

Emergent risks are, by definition, not fully known, or foreseeable, hence effective risk management approaches should rely on extending basic risk management principles to AI risks.

We outline below the three principles of a practical AI risk management framework to help organizations accelerate their AI adoption whilst also effectively meeting new regulatory obligations around AI.

1. Extend traditional cyber security controls to AI systems

- Expand your organization's attack surface to include AI systems.

- Account for the complex nature of the AI system's attack surface and the additional threats and attack vectors they introduce.

- Consider third-party and supply-chain risks introduced through the use of AI systems especially those provided by vendors.

- Extend your incident response processes to cover incidents from attacks on AI systems and any issues arising from the use of AI outputs.

- Understand and train your staff on threats to/from AI systems and their use in your organization.

- Evaluate the applicability of traditional cyber security controls to AI systems. E.g., Infrastructure and cloud controls would be similar.

- Establish comprehensive logging and monitoring for AI systems to effectively meet new regulatory obligations.

2. Expand data governance to manage AI data use

- Understand your regulatory obligations, especially those relate to monitoring and reporting on data used to train AI models.

- Extend your data governance process to control the data allowed to be used for training AI models and the data provided to the AI system post deployment.

- Establish oversight for managing data quality for data sets used in training and ensure logging of data provenance for audit purposes.

- Extend your enterprise data architecture to cover AI data use and ensure data catalogues are upto date.

- Consider data sovereignty obligations when storing data for AI use. E.g., Data sets used in training AI models need to be inclusive whilst ensuring data sovereignty.

- Evaluate your organization's content management and abuse policies concerning use in AI systems, specifically for Gen AI solutions that can generate a wide variety of content on demand.

3. Incorporate AI risk management into existing business processes

- Update existing business processes to consider risks introduced through the use of AI.

- Understand and train your staff on AI risks and how they impact existing risk frameworks.

- Maintain an inventory of all AI models and their associated risk profiles, and continuously assess if their use is commensurate with their risk profile. E.g., The use of Gen AI for document summarization for human consumption may be within the risk tolerance threshold. However, the use of the same model for customer communication without human supervision may exceed the threshold.

- Establish a shared responsibility model for AI use among the developers of the system, teams that deploy and manage the system, and the users of the system. Consider your regulatory obligations when establishing the shared responsibility model as they may place additional restrictions. E.g., The US Executive Order restricts who can develop or use certain types of high-risk AI systems.

Risk Management Approach

Rapid progress in the capabilities of AI systems has introduced some emergent risks that organizations need to consider before they adopt these systems at scale.

Regulators have started to define compliance obligations for organizations that use AI systems, and the failure to implement effective risk management for AI risks will expose organizations to regulatory sanctions.

Also, given the nature of AI risks, some of the controls for risk management may not even be enforceable post-deployment.

Organizations should thoughtfully evaluate the emerging risk categories delineated in this document and ascertain their relevance to their specific organizational context, particularly in light of applicable regulatory obligations.

The risk management methodology presented in this paper can be utilized by organizations to delineate, execute, and oversee their AI risk management strategies.

About letsbloom

letsbloom provides **continuous cloud compliance** to build and manage applications and data on any public cloud. The platform is modular, multi-cloud enabled and acts like the OS for cloud compliance **(DevSecRegOps)**.

Regulatory compliance considerations are the biggest barrier in cloud and AI adoption that limits the FI's ability to innovate with their clients and third parties. letsbloom enables cross-team collaboration to effectively control and govern cloud compliance to accelerate the go-to-market of new products and client experiences.

We are helping Financial Institutions and FinTechs to:

1. Accelerate cloud adoption (Azure/GCP/AWS) in a secure and compliant manner through standardized patterns (infrastructure/ policy/ complia nee-as-code).
2. Leverage cloud AI services including Generative AI services through our **Secure Data Lab app.**
3. Evidence compliance by verifying controls across all cloud assets.
4. View and continuously manage the compliance posture of all cloud assets against regulatory obligations (e.g. NYDFS, MAS, PRA, etc.), industry standards (e.g. CIS, NIST) and internal governance requirements.