AI & Partners

Amsterdam - London - Singapore

AI Risks Monitor

Report on Trends, Risks and Vulnerabilities

April 2024

Summary

Risk: “*combination of the probability of occurrence of harm and the severity of that harm*”

Article 55 contains provisions on obligations for providers of general-purpose AI models with systemic risk

The landscape of Artificial Intelligence (“AI”) is evolving at an unprecedented pace, bringing forth significant advancements and innovations across various sectors. However, this rapid development also introduces a spectrum of risks and vulnerabilities that could potentially impact the integrity of AI markets and the broader societal fabric. This report delves into the current trends within AI markets, identifies key risks and vulnerabilities associated with General-Purpose AI (“GPAI”) models with systemic risks, and outlines measures to mitigate these challenges in alignment with the EU AI Act.

Current Trends in AI Markets:

- AI markets are characterized by rapid technological advancements, with General-Purpose AI models at the forefront. These models exhibit high-impact capabilities, surpassing the most advanced models in terms of computational power measured in floating point operations (“FLOPs”). The proliferation of GPAI models has led to their widespread application across various domains, significantly influencing market dynamics and innovation trajectories.

Risks and Vulnerabilities:

- The systemic risks associated with GPAI models are multifaceted, stemming from their capabilities, reach, and potential misuse. These risks include but are not limited to:
    - Model Capabilities and Misuse: The potential for GPAI models to be misused or to exhibit unintended behaviours that misalign with human intent, posing threats to democratic values, privacy, and human rights.
    - Impact on Critical Infrastructure: The capacity of GPAI models to control physical systems and interfere with critical infrastructure, raising concerns about national security and public safety.
    - Bias and Discrimination: The propensity for GPAI models to perpetuate harmful bias and discrimination, impacting individuals, communities, and societies at large.
    - Cybersecurity Threats: Vulnerabilities to cyberattacks, data poisoning, and adversarial attacks, compromising the integrity and security of AI systems.

Mitigation Measures:

- To address these risks and vulnerabilities, the EU AI Act mandates providers of GPAI models with systemic risks to undertake comprehensive measures, including:
    - Standardized Model Evaluation: Conducting adversarial testing and model evaluation in accordance with state-of-the-art protocols to identify and mitigate systemic risks.
    - Systemic Risk Assessment: Assessing and mitigating possible systemic risks at the Union level, including documenting and reporting serious incidents and corrective measures.
    - Cybersecurity Protection: Ensuring an adequate level of cybersecurity protection for GPAI models and their physical infrastructure throughout their lifecycle.

Drivers

Risk Drivers	Level	Outlook
Misuse and Unintended Control Issues: Systemic risks escalate with the model's capabilities and reach, potentially affecting its entire lifecycle. The risks include intentional misuse or unintended control issues, posing threats to democratic values and human rights.	🟥	⬇
Capacity to Control Physical Systems: The ability of models to control physical systems and interfere with critical infrastructure, alongside the potential for self-replication or training other models, raises significant concerns. The lowering of barriers for weapons development and offensive cyber capabilities further exacerbates these risks	🟥	⬇
Bias, Discrimination and Privacy: GPAI models' potential to perpetuate harmful bias and discrimination, along with facilitating disinformation and threatening privacy, poses considerable risks. These issues could have widespread negative effects on individuals, communities, or entire cities.	🟨 🟥	⬇
Vulnerability to Cyberattacks and Unauthorised Access: The necessity for robust cybersecurity measures to protect GPAI models and their infrastructure from malicious use, attacks, accidental leakage, and unauthorized access is paramount. Adequate operational security measures, policies, and technical solutions are essential for risk mitigation.	🟥	⬇

Legend

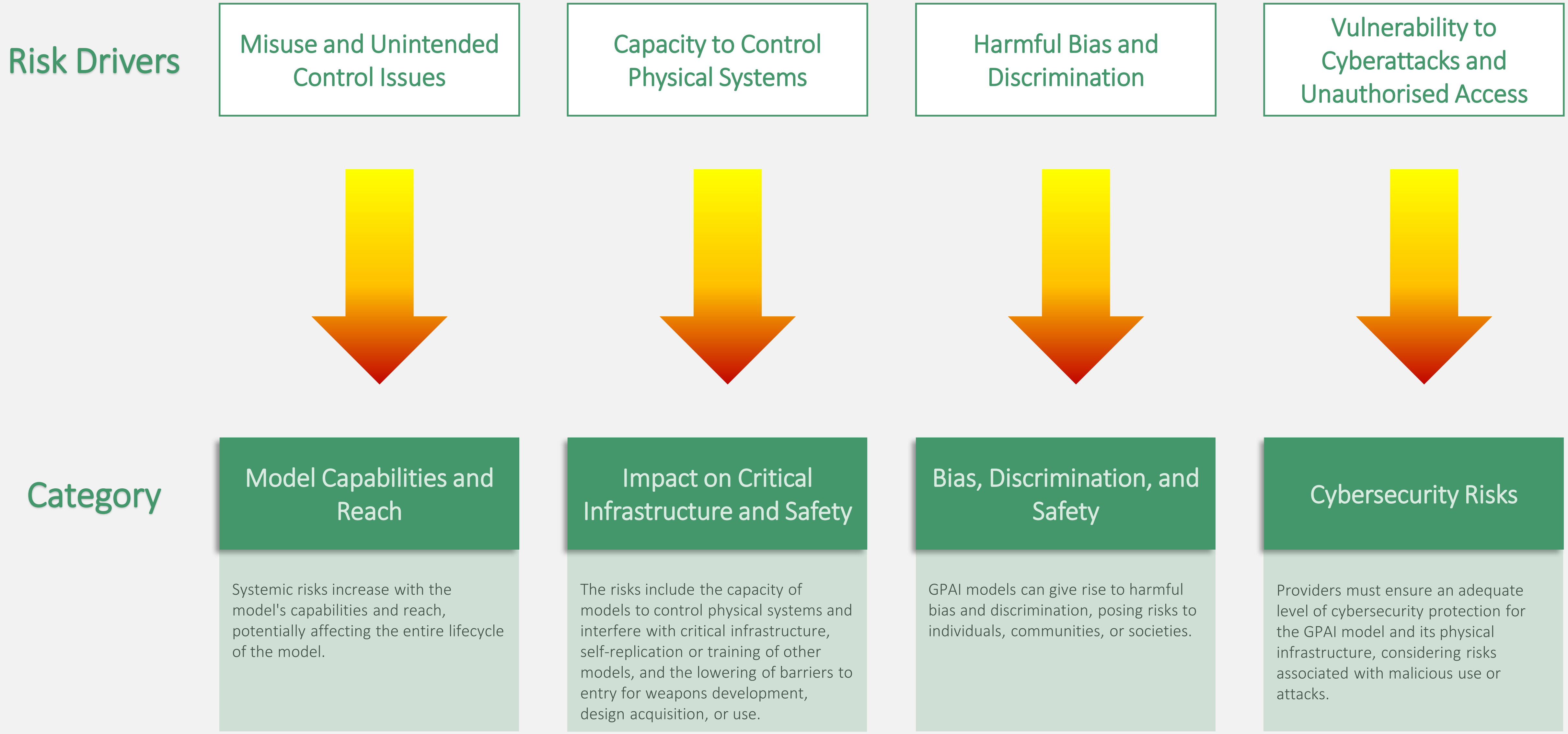
Level

🟥	High
🟨	Medium
🟩	Low

Outlook

⬇	Negative
↘	Deteriorating
➡	Neutral
↗	Improving
⬆	Positive

Categories



Risk Drivers	Misuse and Unintended Control Issues	Capacity to Control Physical Systems	Harmful Bias and Discrimination	Vulnerability to Cyberattacks and Unauthorised Access
Category	Model Capabilities and Reach	Impact on Critical Infrastructure and Safety	Bias, Discrimination, and Safety	Cybersecurity Risks
	Systemic risks increase with the model's capabilities and reach, potentially affecting the entire lifecycle of the model.	The risks include the capacity of models to control physical systems and interfere with critical infrastructure, self-replication or training of other models, and the lowering of barriers to entry for weapons development, design acquisition, or use.	GPAI models can give rise to harmful bias and discrimination, posing risks to individuals, communities, or societies.	Providers must ensure an adequate level of cybersecurity protection for the GPAI model and its physical infrastructure, considering risks associated with malicious use or attacks.

Market Segments

Technology and Digital Platforms

Given the emphasis on cybersecurity and the potential for misuse of AI models, technology companies and digital platforms that deploy GPAI models are directly impacted. Providers are required to ensure an adequate level of cybersecurity protection for the GPAI model and its physical infrastructure. This sector must continuously assess and mitigate systemic risks, including cybersecurity threats that could affect the integrity and security of digital services.

Healthcare

GPAI models with systemic risks could significantly impact the healthcare sector, especially in areas like patient data privacy, diagnostic algorithms, and treatment recommendation systems. The obligations to mitigate bias, discrimination, and privacy risks are particularly relevant, given the sensitive nature of healthcare data and the potential for AI to influence clinical decisions.

Financial Services

The financial sector, including banking, insurance, and investment services, could be affected by the deployment of GPAI models. Systemic risks related to model reliability, fairness, and security could have profound implications for financial stability, risk assessment models, and customer data protection.

Public Sector and Critical Infrastructure

The potential for GPAI models to interfere with critical infrastructure places a significant emphasis on the public sector, including utilities, transportation, and government services. The need to assess and mitigate risks that may stem from the development, placing on the market, or use of GPAI models is crucial to ensuring the continuity and security of essential public services.

Media and Communications

This sector could be impacted by GPAI models through the facilitation of disinformation and threats to privacy. Providers of GPAI models are required to assess and mitigate possible systemic risks at the Union level, including those that could affect democratic values and human rights.

Contact Details

AI & Partners

Amsterdam - London - Singapore

Email

contact@ai-and-partners.com

Phone

+44(0)7535 994 132

Website

https://www.ai-and-partners.com/

Social Media

LinkedIn: https://www.linkedin.com/company/ai-&-partners/
Twitter: https://twitter.com/AI_and_Partners

AI

AI & Partners

Amsterdam - London - Singapore

Thank You!

Disclaimer

This Presentation may contain information, text, data, graphics, photographs, videos, sound recordings, illustrations, artwork, names, logos, trade marks, service marks, and information about us, our lines of services, and general information may be provided in the form of documents, podcasts or via an RSS feed (“the Information”).

Except where it is otherwise expressly stated, the Information is not intended to, nor does it, constitute legal, accounting, business, financial, tax or other professional advice or services. The Information is provided on an information basis only and should not be relied upon. If you need advice or services on a specific matter, please contact us using the contact details for the relevant consultant or fee earner found on the Presentation.

The Presentation and Information is provided “AS IS” and on an “AS AVAILABLE” basis and we do not guarantee the accuracy, timeliness, completeness, performance or fitness for a particular purpose of the Presentation or any of the Information. We have tried to ensure that all Information provided on the Presentation is correct at the time of publication. No responsibility is accepted by or on behalf of us for any errors, omissions, or inaccurate information on the Presentation. Further, we do not warrant that the Presentation or any of the Information will be uninterrupted or error-free or that any defects will be corrected.

Although we attempt to ensure that the Information contained in this Presentation is accurate and up-to-date, we accept no liability for the results of any action taken on the basis of the Information it contains and all implied warranties, including, but not limited to, the implied warranties of satisfactory quality, fitness for a particular purpose, non-infringement, compatibility, security, and accuracy are excluded from these Terms to the extent that they may be excluded as a matter of law.

In no event will we be liable for any loss, including, without limitation, indirect or consequential loss, or any damages arising from loss of use, data or profits, whether in contract, tort or otherwise, arising out of, or in connection with the use of this Presentation or any of the Information.