Toward trusted AI

Amid rapid developments in the AI space – from both a technological and regulatory perspective – the board plays a critical role in enabling progress while managing risks and compliance.

The need for and benefits of trustworthy AI[i]

The accelerated worldwide proliferation of generative AI over the past year inspired business leaders to evaluate use cases to grasp the opportunities of this promising technology in their business. At the same time, predictive AI and machine learning in general also find themselves increasingly the spotlight. According to KPMG's global tech report 2023, organizations consider AI and machine learning the most important technologies for achieving their short-term ambitions.

On the flip side of these exciting opportunities are increased risks around the appropriate and ethical application, development and distribution of AI due to its broader availability and use. Deep fakes facilitate impersonation scams, the distribution of deceptive messages and misinformation. Unexpected behavior by the system can lead to infringement of privacy, copyright and intellectual property rights or a leak of confidential (training) data.

[i] Digital Trust is the expectation by individuals that digital technologies and services – and the organizations providing them – will protect all stakeholders' interests and uphold societal expectations and values. Source: WEF / KPMG – Earning Digital Trust: Decision-Making for Trustworthy Technologies, 2022

"Three in five people are wary of trusting AI systems."

Further, erroneous decisions and biased or manipulated outputs of AI-based systems may negatively affect the fundamental rights, health and safety of us humans or cause harm to our environment.

These increased risks have not gone unnoticed by the public. According to "Trust in artificial intelligence," a global survey conducted by KPMG, three in five people are wary of trusting AI systems. However, 75% of respondents state that they would be more willing to trust AI when assurance mechanisms are in place to signal ethical and responsible use. These could include monitoring system accuracy and reliability, arranging independent AI system reviews and certifications, and applying standards and codes of conduct.

Clearly, then, gaining the trust of the relevant stakeholders, including customers, employees, business partners, regulators and investors, will be fundamental in order to seize the opportunities of AI.

The role of regulation and the EU AI Act?

In response to the increased risks created by AI, the European Union (EU) has reached a ground-breaking provisional agreement on a comprehensive Artificial Intelligence Act (AI Act) that takes a risk-based approach to ensure that AI systems on the EU market are safe and trustworthy, and respect public rights and values. Expected to become law in 2024, with compliance set to be required by 2025, this legislation — the first of its kind — is anticipated to emerge as the de-facto new global standard for AI regulation.

The law imposes obligations on non-EU businesses as it has an extraterritorial effect: any provider placing AI systems on the market or putting them into service within the EU, regardless of its location, falls under the AI Act. So, too, do providers of AI systems located outside the EU, whose system output can or is intended for use in the EU. (See Decoding the EU AI Act for further details.) The consequences of non-compliance might range from restricting market access to significant fines up to EUR 35m or 7% of global turnover.

The EU adopted a broad definition of an AI system: "AI system" means a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

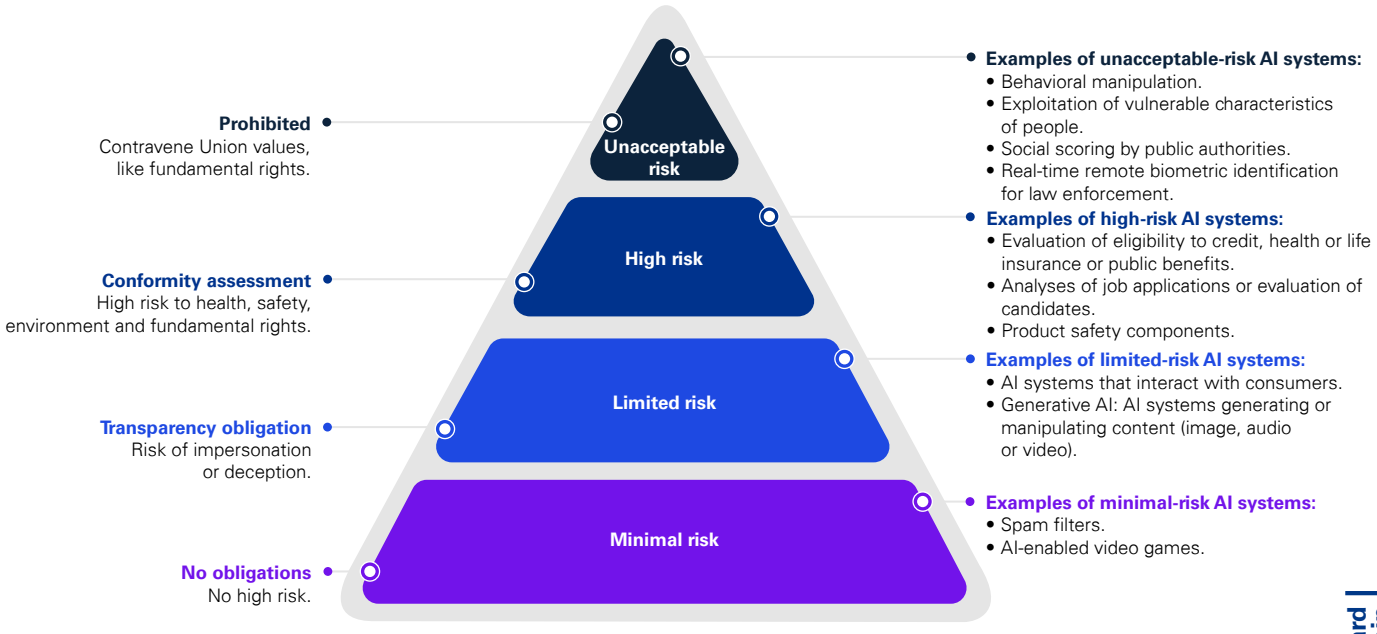
The AI Act defines a risk framework with four categories (see graphic below).

Depending on the category, different requirements for safeguards apply.

- AI systems that fall into the "unacceptable risk" category are prohibited.
- For "high-risk" AI systems a stringent set of requirements applies, covering human oversight, comprehensive risk and impact assessments, data governance practices, incident reporting logging, monitoring and record keeping as well as measures to ensure accuracy, robustness and cybersecurity. Furthermore, such systems are subject to conformity assessment procedures and must affix the CE* mark confirming European conformity.
- Obligations for "limited-risk" AI systems focus on transparency, i.e., that users of the system are informed that they interact with an AI system or that content (images, audio or video) has been created or manipulated by such a system.
- No obligations apply to "minimal-risk" AI systems.

* The CE marking guarantees that the labeled products can be traded without restriction within the EU (or EEA) and ensures uniform protection for consumers within this area in health, safety, and environmental matters. Source: https://www.seco.admin.ch/seco/de/home/Aussenwirtschaftspolitik_Wirtschaftliche_Zusammenarbeit/Wirtschaftsbeziehungen/Technische_Handelshemmnisse/Mutual_Recognition_Agreement_MRA0/CE-Kennzeichnung.html

Risk Framework (AI Act)



The AI Act defines an additional category to cover "General Purpose AI" (GPAI). Large foundation models such as OpenAI's GPT-4 or Google's Gemini fall into this category. Comprehensive obligations – similar to the "high-risk" category – apply.

The regulatory situation in Switzerland

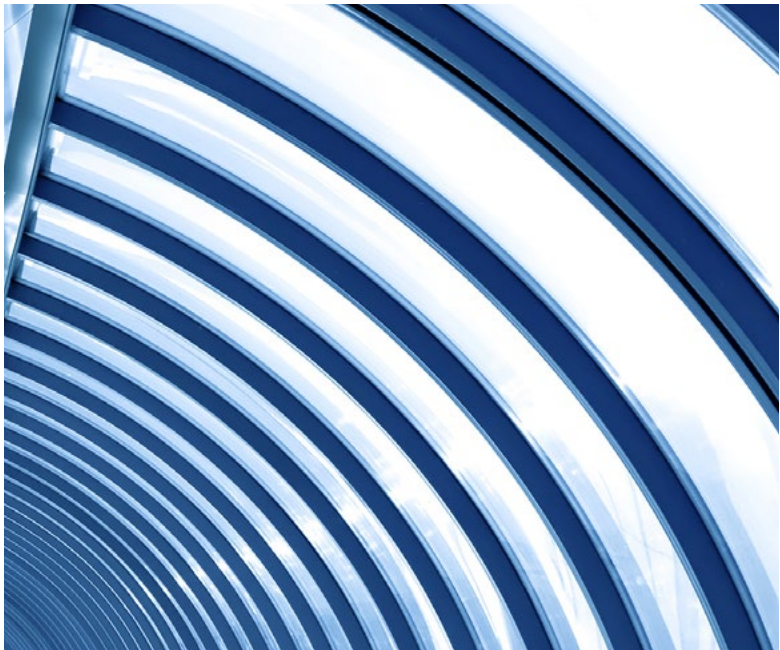
In Switzerland, the Federal Council instructed the Federal Department of the Environment, Transport, Energy and Communications (DETEC) to prepare an overview of possible regulatory approaches to AI, which will build on existing Swiss law and identify possible regulatory approaches for Switzerland that are compatible with the EU AI Act and the Council of Europe's AI Convention. The analysis is expected to be available by the end of this year.

In its Risk Monitor 2023, the Swiss Financial Market Supervisory Authority (FINMA) recognized AI as a strategic risk, particularly with regard to the responsibility for AI decisions, the reliability of AI applications, the transparency and explainability of AI decisions and the equal treatment of financial market clients. FINMA stated that it will monitor the use of AI by supervised institutions.

Key actions for the board

Fueled by the hype around AI technologies, the fear of missing out and employee demand, many organizations are under a very high pressure to deploy AI systems quickly and make them available to a very broad user base. Furthermore, many have been using AI systems for some years. The raising regulatory and public scrutiny means that new requirements may apply to them, and risks of non-compliance increase significantly.

The exceptional dynamic that comes with the current hype around AI systems and related risks pose a major challenge to non-executive boards who are accountable for oversight and risk management.



In the short-term, the board should

1) Ensure that an appropriate governance over AI systems is in place, which includes:

   a. A policy that defines risk categories (aligned with the EU AI Act categories) to cover all your AI systems.
   b. Transparent communication with all stakeholders, including employees, customers and business partners about the use of AI systems and how the organization addresses the related risks.
   c. Implementation or improvement of your AI management system, considering standards such as ISO 42001 and other good practices; your organization may be able to build on other management systems related to information security or privacy that have been implemented in the past.
   d. Employee training on AI ethics and compliance.

2) Ensure that the organization manages AI risks appropriately and adequately reports them to the board, which involves:

   a. Understanding of the risks of AI systems to your organization and stakeholders, including (prospective) employees, (prospective) customers, the public and the entire ecosystem.
   b. A current inventory and classification of your AI landscape.
   c. A gap analysis against applicable regulations, standards and good practices to develop an action plan or roadmap addressing these gaps.
   d. Thorough testing of AI systems to ensure they operate as intended, considering the robustness, fairness and accuracy of the model; the quality of the (training) data; and safety, security and privacy.
      A multidisciplinary team ("red team") may be required to perform these tests with the required quality.
   e. Assessment of risks from AI systems provided or developed by third parties.

In the mid to long term, organizations may strive for a "trusted AI by design" approach to embed trusted AI principles across the entire lifecycle of their AI systems, products and services. Such organizations will be well placed to leverage this trust to differentiate themselves from peers and gain a competitive advantage.

Matthias Bossardt
Partner, Head of Cyber & Digital Risk Consulting
KPMG Switzerland

+41 58 249 36 98
mbossardt@kpmg.com