Objective Assessment of Reasonable Machines? Role and Limitations of Risk Management in the European AI Regulation Efforts*

Richard Skalt
Digital Society Institute
ESMT Berlin
Berlin, Germany
richard.skalt@esmt.org

Nils Brinker†
Digital Society Institute
ESMT Berlin
Berlin, Germany
nils.brinker@esmt.org

Helene Pleil
Digital Society Institute
ESMT Berlin
Berlin, Germany
Helene.pleil@esmt.org

KEYWORDS

Artificial Intelligence (AI), AI Regulation, Risk-Based Approach, EU AI Act, Global Governance, Risk Management

Citation format:

Nils Brinker, Richard Skalt and Helene Pleil. 2023. Objective assessment of reasonable machines? Role and limitations of risk management in the European AI regulation efforts.

1 Introduction

Emerging technologies have an incredible potential to improve citizens quality of life, access to services and contribute to economic growth and substantial increases in productivity. As machine learning capabilities improve and receive more public attention, the downsides of the unregulated use of Artificial Intelligence (AI) are increasingly becoming clear. While AI may be harnessed to accomplish tasks more effectively, it is increasingly impacting human rights, privacy, and most aspects of our digitalized societies. Technologies are not neutral, and nor are the means and motivations by which they are developed (Hare 2022). Great power competition in a race for AI dominance is likely to reshuffle the global balance of power in favor of those countries capable at setting standards for it (Peterson & Hofmann 2022). These standards will be the basis for regulatory frameworks followed by countries, corporations and international organizations and will reflect the values of the countries who contributed to their development, such as fairness, accountability, transparency and non-discrimination (Vanberghen & Vanberghen 2021).

The implementation of these standards comes with challenges for policymakers and global stakeholders hoping to harness the potential of AI without risking the misuse of this technology. At the national level, many countries are adopting or developing AI strategies and policies to govern the development and use of AI, many of whom focus on ethics, privacy and the data used to train its algorithms. In an attempt to lead the development of standards and norms on AI, the European Union (EU)'s AI Act was the first landmark legislative proposal to regulate the development and use of AI in EU member states. In many ways the EU acted as a "norm entrepreneur" by setting an example for other countries aiming to design their own AI governance mechanisms (Manners 2002), but it is not the only model that exists on AI governance. Not only liberal democracies are taking the initiative on regulating AI, with China also enacting regulations that focus on maintaining its social order and societal morality (Sheehan 2023), which are at odds with the liberal democratic principles.

2 The EU's Approach to AI Governance

The EU's approach to AI aims to ensure the responsible use of AI technologies by focusing on safety and fundamental rights, and to strengthen Europe's potential to compete globally by boosting research and industrial capacity. It outlines a regulatory framework with a focus on high-risk AI applications such as healthcare and law enforcement, requiring strict compliance with transparency, accountability and safety standards which are assessed with the help of a dedicated risk assessment process. With the help of risk assessments, the EU aims to address the impact of AI on fundamental rights, such as data privacy and non-discrimination, in line with the EU's commitment to uphold ethical and human-centered AI (European Commission 2023). In contrast to the "context-specific, principles-based approach" favored by the UK (Science and Technology Committee, House of Commons 2023), the EU AI Act is taking a risk-based approach, involving "[…] obligations for providers and those deploying AI systems depending on the level of risk the AI can generate" (European Parliament 2023). As such it has been received by commentators in the UK as " […] a very centralized approach, with some very hard line regulation [… that] does not allow the opportunity for much flexibility", and criticized for "[…] not [being] particularly future-proofed, because they have a static list of high-risk applications that will be under particular regulatory scrutiny […] and would need updating very, very regularly." (Science and Technology Committee, House of Commons 2023).

While the EU and the U.S. are more conceptually aligned on pursuing a risk-based approach, promoting key principles of trustworthy AI and placing emphasis on developing

international standards (not in small part thanks to the EU-U.S. Trade and Technology Council working groups), in practice, there are many differences between their respective AI risk management regimes (Engler 2023). The U.S. approach to AI risk management is highly distributed across federal agencies, often without relying on new legal authorities. Instead, it is accompanied by investment in non-regulatory infrastructure, such as the use of non-binding frameworks on AI, evaluations of facial recognition software, as well as extensive funding of AI research (White House 2023).

3 Risk Assessment in the AI Act

The implementation of a risk-based approach described in the latest proposal for an EU AI Act functions by categorizing AI systems into three different categories, which can also be interpreted as risk categories (Figure 1). Systems falling into each category will then be subject to different types of legal obligations and requirements, which should be proportionate to the risk posed by the use of the AI solution (Floridi et al. 2022). What determines the category is the actual use case or purpose for which an AI system is being used (or can foreseeably be misused).

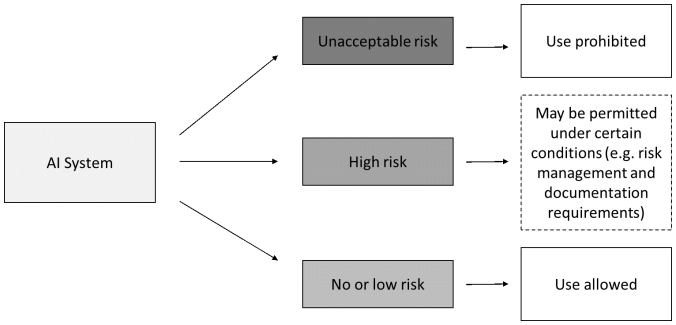

Figure 1: Risk Categories for AI use cases under the AI Act (Floridi et al. 2022)

The first category consists of AI systems that are categorically prohibited (with some caveats for use in law enforcement). The second category consists of so-called high risk AI systems, which are not prohibited per se, but require a number of additional measures to mitigate or minimize the risks associated with their use. The third category, which is not explicitly named, consists of all AI systems that do not fall into either of the first two categories and can therefore be referred to as low-risk AI systems. In general, these AI systems can be freely marketed. However, there are obligations, in particular transparency obligations (see Art. 52 AI Act). They may also be subject to obligations under other regulations, such as the GDPR (see Art. 22 GDPR).

Accordingly, AI systems with an unacceptable level of risk to EU citizens' safety would be prohibited under the AI Act, e.g. systems used for social scoring. The list of banned uses of AI includes (European Parliament 2023):

(1) remote biometric identification systems in publicly accessible spaces;

(2) "post" remote biometric identification systems (with the exception of law enforcement for the prosecution of serious crimes and only after judicial authorization);

(3) biometric categorization systems using sensitive characteristics (e.g. gender, race, ethnicity, citizenship status, religion, political orientation);

(4) predictive policing systems (based on profiling, location or past criminal behavior);

(5) emotion recognition systems in law enforcement, border management, the workplace, and educational institutions; and

(6) untargeted scraping of facial images from the internet or CCTV footage to create facial recognition databases.

Risk (and quality) management requirements play a role in the admissibility and operation of prohibited and high-risk AI systems. In the first case, although AI systems of the first category are generally prohibited, there are use cases that could allow the use of real-time biometric identification systems in public spaces (further elaborated in section I). This is the case when these systems are used to search for potential targets of crime, including missing children, to prevent serious harm to a person (e.g. a terrorist attack), as well as for the prosecution of serious crimes (those punishable by a maximum sentence of at least three years). For AI systems identified as having high risks, quality and risk management methods are required in order to place these systems on the market (further elaborated in section II).

3.1 Prohibited AI Systems

The AI Act does not formulate fully articulated and formalized risk management for the use of generally prohibited AI-systems. However, requirements for the use of otherwise prohibited AI-systems, namely real-time biometric identification in public spaces, also contain criteria to weigh affected rights and interests, as well as criteria to minimize the impact of their use (see Art. 5 (2) AI-Act-P). Since all of the potential use cases include the use of AI by authorities, the means to evaluate its proportionality are (or at least resemble) methods well established in public law (Guggenberger 2019). While those partly inherit comparable weaknesses, as in the risk management process conducted by private entities, those weaknesses are not elaborated in detail within this paper.

There is, however, another important aspect that must be considered: While the use of some AI-systems may be allowed for specific use cases that are deemed acceptable, it neglects the characteristic of AI as a universal tool. Once the technical infrastructure is implemented, the question of the use for other purposes becomes a legal question, not one of technical possibility. Technically, it is no different to use biometric

identification to surveil political dissidents or search for missing children. Therefore, a technical implementation with only legal barriers for human-rights-violating use cases requires a functioning constitutional state.

3.2 High Risk AI Systems

In the domain of high-risk AI systems, the AI Act works as a product safety regulation (Roos and Weitz 2021). Providers of AI-systems must fulfill certain obligations to mitigate or reduce the potential harm of the AI systems they bring to market (Art. 16ff. AI-Act-P). One part of those obligations is the introduction of risk management, while comparable requirements can also come into effect through harmonized norms or common specifications (see e.g. conformity assessment of medical products, Annex 1, chap. 1 Nr.3 Medical Devices Directive).

The risk management required by the AI-Act consists of three steps (Figure 2):

(1) **Risk Identification** (Art. 9 (2) a AI-Act-P): Risks estimated to occur through the intended use and foreseeable misuse of the AI System (Art. 9 (2) b AI-Act-P)

(2) **Risk Evaluation**:

A. Evaluation of risks estimated to occur through the intended use and foreseeable misuse of the AI System (Art. 9 (2) b AI-Act-P)

B. Evaluation of risks identified through operational monitored data (Art. 9 (2) c AI-Act-P))

(3) **Risk Treatment**: (Art. 9 (2) d AI-Act-P)

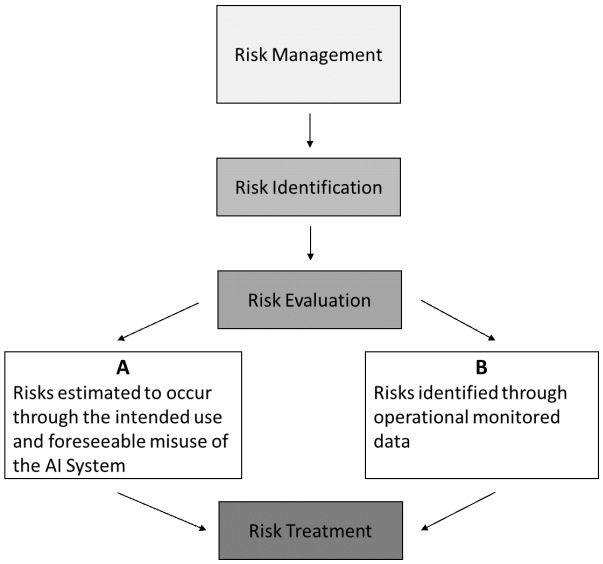

Figure 2: Risk Management process as required in the EU AI Act

For the regulator, risk management fulfills a double purpose: First, one requirement of the result of the risk management process is that risks are to be eliminated or reduced to a level that is accumulated acceptable. The compliant use of AI systems is, therefore, not possible. Hence, risk management is a part of the evaluation of the overall admissibility of an AI system. The second function is to minimize the operational risk associated with the AI system. As an iterative process, risk management aims for continuous optimization and not just for an initial assessment (Art. 9 (2) AI-Act-P).

The technical and organizational measures chosen in step three should be proportionate to their use of risk reduction concerning the effort of their implementation (Art. 9 (3) AI-Act-P). The Addressee of the requirement to conduct the risk assessment is the provider of an AI system (Art. 16 (a) AI-Act-P); therefore, the legal person that markets an AI system or puts it into service. Also, a legal person that trades an AI under their name changes the purpose of the AI system or conducts technical modifications can be assigned with obligations of the provider (Art. 28 (a) AI-Act-P). In every case, the entity that must conduct the risk management has a self-interest in the actual use of the AI.

3.3 Disadvantages of Regulation based on Risk Management

In general, risk-based approaches were seen as progress compared to "one size fits all" type of regulations. Regulations with strict criteria that have to be met in order for certain requirements or legal consequences to come into effect are, in some domains, seen as too static to do justice to dynamic and complex real-life scenarios (Schröder 2019). Also, risk management can be seen as a way to dynamically adjust the efforts an addressee of certain laws has to put into compliance measures and to put those into proportion. Also, a risk-based approach and risk management as part of it, in theory, seems to be a way to balance the protection against the potential harm of AI and to give necessary room for innovation (cf. Hammon et al. 2023).

Risk management is not a process with a clear and deterministic outcome. Each step in the process leaves room for error, interpretation, and individual judgment and preference. This begins with the risk identification step, which enumerates the potential harms of an AI system. While risk management generally requires an objective inventory, the subjective view of the entity performing the risk management process cannot be eliminated. This is also true for the following risk management steps: risk assessment and implementation of mitigation measures. The AI Act attempts to minimize such subjective factors by explicitly stating the factors to be considered in risk management (Art. 9 AI-Act-P).

This subjectivity is especially true when it comes to intangible damages, which cannot be definitively quantified. Even if qualitative assessments should be as objective as possible, they cannot be mathematically precise but are based on verbal arguments. Even if these verbal arguments can be assigned numerical values, this assignment is based on subjective associations. Therefore, it is possible that the result of a risk

assessment is not a matter of facts but of rhetoric. The subjectivity of the risk management process is not a new phenomenon (see, e.g., Ramnarine 2015). However, it must be considered when using risk management as a regulatory tool. It helps to look at the roots of risk management not as a definitive tool to algorithmically calculate future steps but as a means (only) to facilitate decision-making.

Arguing in favor of one's own interests in risk management is not necessarily a sign of bad faith. On the one hand, it is natural and rational for an entity to act in its own best interests. Moreover, the assessment of third-party risks, especially immaterial ones, is not an easy task for entities whose core competencies are in other domains (such as engineering). Therefore, the entities legally required to conduct the risk assessment may simply lack the competence to conduct, for example, a complex human rights assessment of their product.

The AI Act recognizes that risk management as a simple means of self-governance may not have the desired effect. For this reason, the AI Act introduces a system of authorities to ensure that risk management is carried out in accordance with the law (see conformity assessment of the AI Act, Art. 19 AI-Act-P).

This accountability of the entities performing risk management at least increases the objectivity of the risk assessment. However, effective compliance enforcement in this way is costly and may be slow to take effect. The same applies to the formulation of common specifications (Art. 41 AI-Act-P) or harmonized standards (Art. 40 AI-Act-P). In specific cases, these criteria may be effective in minimizing objectivity in the risk management process, especially in the step of implementing mitigation measures. However, the formulation of such common criteria is an even slower process.

4 Lessons to be Learned

In summary, the EU AI Act serves as a valuable model for the UN and other nations in developing global AI governance principles. The weaknesses of risk management outlined above are not severe enough to qualify the process as an inappropriate tool in general. Nevertheless, the weaknesses, particularly its subjectivity and its tendency to downplay the risks posed by third parties, should be taken into account. When applying risk management to the lawful use of high-risk AI, careful consideration must be given to which use cases are categorized as high-risk and which are prohibited from the outset. For example, the AI Act categorizes the use of polygraphs as a tool in asylum and border control management as high risk and consequently does not categorically prohibit it. It remains questionable whether any risk management carried out will effectively take into account the risks to people in a critically vulnerable position.

By considering the lessons to be learned from the EU's regulatory efforts, the UN can work towards establishing a comprehensive, ethical, and collaborative framework that addresses the unique challenges posed by AI on a global scale. It should be noted, however, that all EU member states share the same fundamental values as well as similar legislative and political systems. The same cannot be said for UN member states, which will be a major challenge given that AI regulations are also a reflection of values.

ACKNOWLEDGMENTS

We would like to acknowledge the support of our colleagues at the Digital Society Institute at ESMT Berlin in supporting us with brainstorming for this paper.

REFERENCES

[1] Engler, A. (2023). The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. Brookings Institution. https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/

[2] European Commission (2023). A European approach to artificial intelligence. Shaping Europe's digital future. Digital Strategy. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

[3] European Parliament (2023). MEPs ready to negotiate first-ever rules for safe and transparent AI. Press release. https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai

[4] Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). CapAI-A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. Available at SSRN 4064091.

[5] Guggenberger, L. (2019). Einsatz künstlicher Intelligenz in der Verwaltung. In: Neue Zeitschrift für Verwaltungsrecht, p. 844-850.

[6] Hare, S. (2022). Technology is not neutral: A short guide to technology ethics. London Publishing Partnership.

[7] House of Commons (2023). The governance of artificial intelligence: interim report. Chapter 6: The international dimension. Science and Technology Committee. https://committees.parliament.uk/publications/41130/documents/200993/default/#1769%20illustrated%20front%20cover.indd%3A.10952%3A2087

[8] Hammon, C., Buyers, J., Docquier, B., Schefzig, J., Vickery, K., Sharpe, T., Kirschke-Biller, J. (2023). Is the proposed European AI Act innovation-friendly?, https://explore.osborneclarke.com/tmcannualreview2023/is-the-proposed-european-ai-act-innovation-friendly.html

[9] Manners, I. (2002): Normative Power Europe: A Contradiction in Terms? In: Journal of Common Market Studies 40 (2), p. 235-258.

[10] Peterson, D. & Hoffman, S. (2022): Geopolitical Implications of AI and Digital Surveillance Adoption. In: Foreign Policy at Brookings.

[11] Ramnarine, E. (2015). Understanding Problems of Subjectivity and Uncertainty in Quality Risk Management. In: Journal of Validation Technology, 21(4).

[12] Roos, P., Weitz, C. A. (2022). Hochrisiko-KI-Systeme im Kommissionsentwurf für eine KI-Verordnung. In Multimedia und Recht, p. 844-851

[13] Sheehan, M. (2023). China's AI Regulations and How They Get Made. Carnegie Endowment. https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117

[14] Vanberghen, C. & Vanberghen, A. (2021). AI Governance as a Patchwork: The Regulatory and Geopolitical Approach of AI at International and European Level. In: Synodinou, TE., Jougleux, P., Markou, C., Prastitou-Merdi, T. (eds) EU Internet Law in the Digital Single Market. Springer, Cham. https://doi.org/10.1007/978-3-030-69583-5_9

[15] Weizenbaum, J. (1976), Computer Power and Human Reason - From Judgement to Calculation. W.H. Freeman and Company

[16] White House (2022). Biden-Harris Administration Takes New Steps to Advance Responsible Artificial Intelligence Research, Development, and Deployment. Fact Sheet. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/23/fact-sheet-biden-harris-administration-takes-new-steps-to-advance-responsible-artificial-intelligence-research-development-and-deployment/