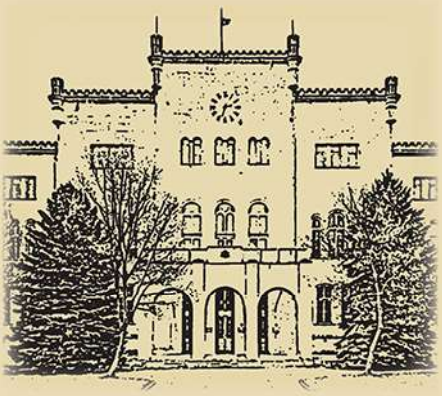
СЪВРЕМЕННИ АСПЕКТИ НА СИГУРНОСТТА

ПРЕДИЗВИКАТЕЛСТВА, ПОДХОДИ, РЕШЕНИЯ

Годишна студентска научна сесия

ФАКУЛТЕТ КОМАНДНО-ЩАБЕН

27 септември 2024 г.



ВОЕННА АКАДЕМИЯ „ГЕОРГИ СТОЙКОВ РАКОВСКИ“

ETHICAL & LEGAL RESPONSIBILITY FOR ARTIFICIAL INTELLIGENCE. THE REGULATORY REGIME FOR HIGH-RISK ARTIFICIAL INTELLIGENCE SYSTEMS

Polina Petrova, Rada Stoilova
Law and Internet Foundation

Abstract: With due regard to nowadays' societies confronting harsh political or environmental threats, the protection of critical infrastructure has become both a challenge and a priority within the European Union. The TESTUDO project ("Autonomous Swarm of Heterogeneous resources in infrastructure protection via threat prediction and prevention"), financed by the "Horizon Europe" programme accentuates on these needs by providing innovative solutions for continuous monitoring, prevention and prediction of hazardous events. The project applies state-of-the-art technologies and AI-based models to enhance cybersecurity, as well as ensures continuous surveillance of the environment and autonomous resource allocation. The TESTUDO platform being developed within the project includes the latest technologies for: fast and coordinated automated response; robust communication networks with highest coverage; cognition capabilities for detection of threats; increased situational awareness and decision-making; exploiting novel HMI and XR technologies.

In view of the above, an interdisciplinary approach is being applied in the current paper, discussing applicable ethical and legal aspects that arise in the context of Artificial Intelligence

(hereinafter "AI"), thereby allowing for the alignment of AI systems with human well-being, respect for human autonomy, privacy, social responsibility, transparency, security, etc. The purpose of the current report is thus to highlight that the minimization of potential adverse effects with regard to the aforementioned technical aspects is of primary relevance, by also explaining how the said objectives are to be achieved.

***Keywords**:* Artificial Intelligence; AI; AI Act; Critical Infrastructure; High-Risk AI Systems; Cybersecurity; Ethics; Law; TESTUDO

1. Introduction

With due regard to nowadays' societies confronting harsh political or environmental threats, the protection of critical infrastructure has become both a challenge and priority within the European Union (hereinafter "EU"). The TESTUDO project ("Autonomous Swarm of Heterogeneous resources in infrastructure protection via threat prediction and prevention"), financed by the "Horizon Europe" programme, accentuates on these needs by providing innovative solutions for continuous monitoring, prevention and prediction of hazardous events. The project applies state-of-the-art technologies and AI-based models to enhance cybersecurity, as well as ensures continuous surveillance of the environment and autonomous resource allocation. The TESTUDO platform being developed within the project includes the latest technologies for: fast and coordinated automated response; robust communication networks with highest coverage; cognition capabilities for detection of threats; increased situational awareness and decision-making; exploiting novel HMI and XR technologies. The pilots applying and testing the technical developments are three and are as follows: 1. Disruptive online events in water reservoirs; 2. Chemical fire in tunnel provoked by an electric vehicle; 3. Synchronized attack on water treatment facilities.

In view of the above, an interdisciplinary approach is being applied in the current paper, discussing applicable ethical and legal aspects that arise in the context of Artificial Intelligence (hereinafter "AI"), thereby allowing for the alignment of AI systems with human well-being, respect for human autonomy, privacy, social responsibility, transparency, security, etc. The purpose of the current report is thus to highlight that the minimization of potential adverse effects when technological solutions similar to the aforementioned context are being developed and deployed, is of primary relevance, by also explaining how

the said objectives are to be achieved. Therefore, the current paper is preoccupied with bringing to the forefront the bedrock principles surrounding AI Ethics, the essence of the regulatory regime under the AI Act, eventually delving into the specificities of the most regulated type of AI systems, namely "High-Risk AI Systems".

The applied research method is qualitative, the relied upon data is secondary, and the narrative is an amalgam of both descriptive and analytical reasoning. The forthcoming text relies heavily on the applicable ethical & legal framework related to AI, thereby constructing a narrative around both the relevance and the implementation of a risk-management approach.

2. Exposure

The forthcoming passages review the implications of the sources which must be not a mere constituent part but, among others, the central pillar, and the very core, of any compliance monitoring process concerning AI systems, namely:

– Regulation (EU) 2024/1689 […] (hereinafter "AI Act" and "AIA");

– Ethics guidelines for trustworthy AI (hereinafter "Guidelines");

– The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment (hereinafter "ALTAI").

Prior to substantively overviewing the applicable provisions, it is worth acknowledging the official definition of an AI system, as defined by Article 3(1) of AIA, namely:

"'AI system' means a machine-based system that is 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." (Regulation (EU) 2024/1689 of the European Parliament and of the Council […], 2024, Art. 3(1)).

2.1. *Ethical AI*

Constituting e set of moral principles, ethics assist in discerning between right and wrong, thereby serving as guidelines for best practice. Even though ethics is not capable of achieving what regulation does, namely, to codify end enforce ethically desirable behavior, meaning it does not constitute a primary source of law, the current section begins with ethical concepts, and not with the AI Act itself, as the former underpin much of the logic behind the regulatory

efforts surrounding the latter. For example, the Ethics Guidelines for Trustworthy AI (April 2019), the Assessment List for Trustworthy Artificial Intelligence (ALTAI) (end of 2020), as well as the White Paper on Artificial Intelligence (February 2020), have all preceded the AI Act (August 2024). It is thus worth commencing with the bedrock foundations of AI ethics, and only then delving into the regulatory side of AI.

2.1.1. *Ethics Guidelines and Trustworthy AI Self-Assessment*

First and foremost, as postulated by the Ethics Guidelines for Trustworthy AI, trustworthy AI should be as follows: lawful, meaning it respects all applicable laws and regulations; ethical, meaning it respects ethical principles and values; robust, in the context of technical aspects, while also paying due regard to social environment considerations. Stemming from these principles, 7 key requirements have been put forward by the Guidelines, explaining further what conditions must be met by AI systems (European Commission 2019).

- Human agency and oversight: AI systems should foster the enjoyment of fundamental rights, including, to allow for human beings to take informed decisions. Human-in-the-loop, human-on-the-loop, and human-in-command approaches need to be present, thereby ensuring proper oversight.
- Technical Robustness and safety: Resilience and security need to be ensured. To secure the presence of a safety mechanism, a fallback plan must be in place. Accuracy, reliability, as well as reproducibility, are also of core relevance in this regard. Unintentional harm can thus be minimized and prevented.
- Privacy and data governance: Adequate data governance mechanisms must be ensured, in addition to the full respect for privacy and data protection. The quality and integrity of data, including the guarantee of legitimate access to data, should also be assured.
- Transparency: Transparency should be applied to the data, system and AI business models, and traceability mechanisms may assist in achieving this goal. Moreover, the explanation of the AI systems and their decisions should be adapted to the concerned stakeholder. The fact that an interaction with an AI system is taking place must be well communicated, as well as the capabilities and limitations of the system.
- Diversity, non-discrimination and fairness: Unfair bias must be avoided in the first place. AI systems should not merely be accessible to all, but they should foster the involvement of relevant stakeholders throughout the entire life cycle.
- Societal and environmental well-being: AI systems should be sustainable and environmentally friendly, thereby ensuring that they benefit all human

beings, including future generations. In other words, the environment, including other living beings, should be considered, as well as the social and societal impact that is being made.

- Accountability: Responsibility and accountability for AI systems, including regarding their outcomes, should be ensured through concrete mechanisms, e.g., assessing algorithms, data and design processes through regular audits, particularly in critical applications, is of crucial importance therein. In addition, redress in an adequate and accessible manner should be made available, too.

Furthermore, ALTAI is a practical tool that has been subsequently developed to translate the Ethics Guidelines into a (self-assessment) checklist that developers and deployers of AI may use in their effort to implement the key requirements in practice (European Commission, 2020).

2.2. *Risk Assessment Methodology under the AI Act*

The AI Act is the first legal framework on artificial intelligence worldwide. Aiming at fostering trustworthiness in AI systems, both in Europe and beyond, the AIA is thus ensuring that safety and fundamental rights of people and businesses are guaranteed.

In its chapters I-IV, AIA avails a risk assessment methodology by distinguishing amongst the following types of risk: unacceptable risk – prohibited (e.g., social scoring systems and manipulative AI); high-risk AI system – subject to stringent regulation; limited risk AI system – subject to lighter obligations, compared to high-risk ones (the fact that end-users are interacting with AI, i.e., chatbots and deepfakes, must be made known; this being the obligation of developers and deployers); and, AI system with minimal risk – not regulated, at least at present (e.g., major AI applications currently available on the EU single market, namely, AI enabled video games and spam filters). In addition, General Purpose AI also falls under the scope of AIA. Despite the modular approach in assessing risk, the most major part of the postulated obligations falls on providers, i.e., developers of high-risk AI systems. In other words, on those that intend to place them on the market, or put into service, high-risk AI systems in the EU (including third country providers when their system is being used in the EU). On the other hand, users, i.e., deployers (and not end-users), of high-risk AI systems, located in the EU, as well as third country providers, where the output of the said system is being used in the EU, are also subject to certain obligations, yet less than compared to developers (Regulation (EU) 2024/1689 of the European Parliament and of the Council […], 2024, Chap. I-IV).

2.2.1. High-Risk AI Systems

Article 6 of the AI Act explains that to be defined as high-risk, AI systems shall either: a.) be used as a safety component of a product; b.) are themselves products covered by EU laws as per Annex I, and as such, are required to undergo a conformity assessment by a third-party under the said Annex I laws; or, c.) fall under the scope of the below-referenced use cases listed in Annex III (Regulation (EU) 2024/1689 of the European Parliament and of the Council […], 2024, Art. 6(1) (2)); with an exception applying to: AI systems performing a narrow procedural task; when an improvement concerns the result of a previously completed human activity; cases of detections of decision-making patterns, including in cases of deviations from the same, where the functionalities are not intended to replace or influence the assessment previously completed by human, without adequate further human overview (Regulation (EU) 2024/1689 of the European Parliament and of the Council […], 2024, Art. 6 (3)). High-risk is always present, when profiling of individuals is involved, i.e., automated processing of personal data for the assessment of aspects of personal life, e.g., performance at work, economic situation, health, preferences, interests, reliability, behavior, location or movement (Regulation (EU) 2024/1689 of the European Parliament and of the Council […], 2024, Art. 6).

Article 6 of AIA further postulates that when a provider considers that his or her systems that is otherwise falling under the scope of the below-referenced use cases, believes that it does not fall therein, the same is still obliged to document the assessment in question prior to deployment on the market, or to putting it into service. Article 49 specifies that such provider is expected to take the respective registration actions, as envisaged by Article 71 on EU database for high-risk AI systems listed in Annex III (Regulation (EU) 2024/1689 of the European Parliament and of the Council […], 2024), Art. 6(4).

- Non-banned biometrics: systems using remote biometric identification, except for cases of verifying that a person is the same as claimed, as well as systems using biometric categorization which infers sensitive/protected attributes/characteristics; systems using emotion recognition.
- Critical infrastructure: AI systems to be used as safety components in the management and operation of critical digital infrastructure, road traffic, as well as the supply of water, gas, heating and electricity.
- Education and vocational training: AI systems which determine access, admissions or assignment to educational and vocational training institutions; the evaluation of learning outcomes; processes during which appropriate level

of education for individuals is being assessed; the monitoring and detection of prohibited student behavior during tests.

● Employment, workers management and access to self-employment: AI systems used for targeted job ads, analyzing and filtering applications, evaluating candidates, namely, usage for recruitment purposes; contracts, their promotion and termination; the allocation of tasks based on personality traits or characteristics and behavior, including monitoring and evaluation of performance.

● Access to and enjoyment of essential public and private services: AI systems used by public authorities for the assessment of benefits and services eligibility, their allocation, reduction, revocation, or recovery, etc.; the evaluation and classification of emergency calls, including dispatch prioritizing; risk assessments and pricing in health and life insurance.

● Law enforcement: AI systems used in the assessment of an individual's risk of becoming a crime victim or offending, or re-offending, the latter being not solely based on profiling or assessing personality traits of past criminal behavior; polygraphs; evaluation of evidence reliability during criminal investigations or prosecutions; profiling when criminal detections, investigations, or prosecutions, take place.

● Migration, asylum and border control management: Assessing of irregular migration or health risks; polygraphs; examining application for asylum, visa, residence permits, as well as eligibility complaints related thereby; detection, recognition or identification of individuals, not for travel documents.

● Administration of justice and democratic processes: AI being used in the research and interpretation of facts, including in the application of law to concrete facts, as well as in alternative dispute resolution; influence during elections/referenda and, generally, voting behavior related thereto; outputs not directly interacting with people, e.g., tools used for organization, optimization and structuring political campaigns, being excluded.

Furthermore, concerning "critical infrastructure", in Recital 55, the legislators justify the high-risk classification by explaining that the failure or malfunction of such systems could "put at risk the life and health of persons at large scale and lead to appreciable disruptions in the ordinary conduct of social and economic activities". The said recital further explains that safety components of critical (digital) infrastructure play pivotal role in the protection of the physical integrity of the latter, or the health and safety of persons and property, but that they are not necessary for the given system to function. It is the failure or malfunctioning of such components that could be the direct cause of risks to the physical integrity of the critical infrastructure though, thereby

affecting health and safety of persons and property. On the other hand, AI systems intended to be used solely for cybersecurity purposes in the context of the safety of critical infrastructure, e.g., systems for monitoring water pressure or fire alarm controlling systems in cloud computing centers, do not fall under this categorization (Regulation (EU) 2024/1689 of the European Parliament and of the Council […], 2024, Recital 55).

Key Obligations on Providers

Subject to the most stringent rules under AIA, high-risk AI systems must meet the following additional requirements (Regulation (EU) 2024/1689 of the European Parliament and of the Council […], 2024, Chap. III (Sections 2-3):

– Risk management system established for the entire lifecycle (Art. 9);

– Data governance conduction, ensuring the relevance and representativeness of the training, validation and testing datasets, including their completeness and lack of errors (Art. 10);

– Technical documentation demonstrating compliance and availing the respective information to authorities, enabling assessment thereof (Art. 11);

– Design featuring record-keeping, thereby enabling the automatic recording of events that may identify national level risks and substantial modifications throughout the system's lifecycle (Art. 12);

– Instructions for usage to downstream deployers (Art. 13);

– Design enabling deployers to implement human oversight (Art. 14);

– Design achieving accuracy, robustness, cybersecurity (Art. 15);

– Quality management system established for enabling compliance (Art. 17).

Key Obligations on Deployers

The deployment of high-risk AI systems, including by public authorities and private organizations providing essential services, such as banks, insurers, hospitals and schools, etc., entails specific obligations to ensure responsible usage of the given system. The key aspects related thereto are as follows (Regulation (EU) 2024/1689 of the European Parliament and of the Council […], 2024, Chap. III (Section 3)):

– Fundamental Rights Impact Assessment prior to deploying (Art. 27);

– Human oversight by natural persons who have the necessary competence, training and authority, as well as the necessary support (Art. 26(2));

– Ensuring that input data is relevant and sufficiently representative in the context of the intended purpose of the system (Art. 26(4));

– Suspension of use of the system in case of risks at national level (Art. 26(5));

– Reporting serious incidents immediately, to the provider, then to the importer or distributor, and to respective market surveillance authorities (Art. 26(5));

– Keeping the automatically generated logs (Art. 26(6));

– When the deployer is an employer, the same shall, prior to putting into service the system at the workplace, inform workers' representatives and the affected workers that they are being subject to the use of the said system (Art. 26(7));

– When the deployer is a public authority, compliance with registration obligations referred to in Art. 45 of AIA (Art. 26(8));

– GDPR data protection compliance, including data protection impact assessment under Article 35 of Regulation (EU) 2016/679 (Art. 26(9));

– Informing natural persons that they are being subject to the use of the system (Art. 26(11)).

3. Conclusion

Being the first of its kind, the AI Act is undoubtedly a milestone in the realm of AI, not only within the EU, but globally, too. The potential of new technologies using AI to reshape nowadays' societies and economies has been foreground by AIA. In this regard, the introduction of the multi-layered risk-based approach is evidential for the fact that not all AI systems are on the same ground concerning their societal impact. In this regard, it has been witnessed that within the discussed approach, high-risk AI is the key category on which most obligations are incurred.

Acknowledgements: The preparation of the current paper is funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

NOTES

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and

(EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), 2024. *Official Journal L series*.

2. European Commission, Directorate-General for Communications Networks, Content and Technology, 2019. *Ethics guidelines for trustworthy AI*. Brussels, Belgium: Publications Office of the European Union, pp. 11 – 13.

3. European Commission, Directorate-General for Communications Networks, Content and Technology, 2020. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. Brussels, Belgium: Publications Office of the European Union.

4. European Commission, Directorate-General for Communications Networks, Content and Technology, 2021. *Ethics by Design and Ethics of Use Approaches for Artificial Intelligence*. Brussels, Belgium: Publications Office of the European Union.

5. HENZ, P., 2021. Ethical and legal responsibility for Artificial Intelligence. *Discover Artificial Intelligence, vol. 1*. Springer, p. 4.

6. KOP, M., 2021. EU Artificial Intelligence Act: The European Approach to AI. *SSRN Electronic Journal*.

7. MÖKANDER, J., 2023. *Auditing of AI: Legal, Ethical and Technical Approaches*. NY, USA: Springer.

Polina Petrova
Legal Expert
ORCID iD: 0009-0003-6708-2645
Law and Internet Foundation
Sofia, Bulgaria
E-mail: polina.petrova@netlaw.bg

Rada Stoilova
Legal Expert
ORCID iD: 0009-0005-5730-4499
Law and Internet Foundation
Sofia, Bulgaria
E-mail: rada.stoilova@netlaw.bg