Legal Issues in the Digital Age. 2023. Vol. 4. No. 3.
Вопросы права в цифровую эпоху 2023 Том 4. № 3.

Research article
УДК: 340
DOI:10.17323/2713-2749.2023.3.97.116

Analysing Risk-Based Approach in the Draft EU Artificial Intelligence Act

Dmitryi Leonidovich Kuteynikov[1], Osman Alikovich Izhaev[2]

[1, 2] Social and Legal Approaches to Applying Artificial Intelligence and Robotics Laboratory, State and Law Institute, Tyumen State University, 6 Volodarskogo Str., Tyumen, Russia, 625003

[1] kuteynikov@me.com; https://orcid.org/0000-0003-1448-3085 https://elibrary.ru/ author_profile.asp?id=776358, izhaev.osman@gmail.com, https://orcid.org/0000-0003-3777-8927, https://elibrary.ru/author_profile.asp?id=827391 SPIN-код: 3197-8857, AuthorID: 827391

[2] izhaev.osman@gmail.com, https://orcid.org/0000-0003-3777-8927, https://elibrary.ru/author_profile.asp?id=827391

Abstract

The article delves into the risk-based approach underpinning the draft EU Artificial Intelligence Act. Anticipated to be approved by the end of 2023, this regulation is poised to serve as a cornerstone in the European Union's legal framework for governing the development and deployment of artificial intelligence systems (AI systems). However, the ever-evolving technological landscape continues to present novel challenges to legislators, necessitating ongoing solutions that will span years to come. Moreover, the widespread proliferation of foundation models and general purpose AI systems over the past year underscores the need to refine the initial risk-based approach concept. The study comprehensively examines the inherent issues within the risk-based approach, including the delineation of AI system categories, their classification according to the degree of risk to human rights, and the establishment of optimal legal requirements for each subset of these systems. The research concludes that the construction of a more adaptable normative legal framework mandates differentiation of requirements based on risk levels, as well as across all stages of an AI system's lifecycle and levels of autonomy. The paper also delves into the challenges associated with extending the risk-oriented approach to encompass foundation models and general purpose AI systems, offering distinct analyses for each.

Keywords

artificial intelligence; AI systems; large language models; generative AI systems; foundation models; general purpose AI systems; Draft Artificial Intelligence Act; risk-based approach; conformity assessment procedure; audit of AI systems.

Acknowledgments: the paper is published within the project of supporting the publications of the authors of Russian educational and research organizations in the Higher School of Economics academic publications.

For citation: Kuteinikov D.L., Izhaev O.A. (2023) Analyzing Risk-Based Approach in the Draft EU Artificial Intelligence Act. *Legal Issues in the Digital Age,* vol. 4, no. 3, pp. 97–116. DOI:10.17323/2713-2749.2023.3.97.116

Introduction

The draft EU Artificial Intelligence Act[1] (AIA Draft) is a comprehensive act intended to regulate interactions in most of the areas related to the development and application of AI systems [Veale M. et. al., 2021: 112]. The EU initiated its development in 2018 involving a wide range of experts and the business community. As part of this work, a number of conceptual papers were presented that gradually formalised the key principles on which the future act was based. [2] The first text of the Draft was published in April 2021. In June 2023, the European Parliament approved the document with its amendments. This was followed by the trilogue stage, which involves agreeing on a unified text of the document on the basis of the positions worked out by the agencies. According to Euro MPs, the Draft will be ap-

[1] Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Available at: URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (accessed: 30.08.2023)

[2] The most important of them are: Ethics guidelines for trustworthy AI. Available at URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai; Policy and investment recommendations for trustworthy Artificial Intelligence. Available at URL: https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence; High-Level Expert Group on AI: Final assessment list on trustworthy AI (ALTAI). Available at URL: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment; White Paper On Artificial Intelligence — A European approach to excellence and trust. Available at URL: https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf (accessed: 08.10.2023).

proved by the end of 2023. According to the latest version of the text, the Draft will be in force twenty four months after its approval.

The AIA Draft is risk based, that involves differentiating the requirements for bringing AI systems to market depending on their potential risk to human rights. In one form or another, this approach is the basis of regulatory concepts in many countries, including the USA[3], China[4], and Russia.[5] However, it is in the EU that it is closest to legislative implementation. Legislators in other countries and regions are either closely studying the European experience or directly declare their desire to adopt it [Gstrein O., 2022: 755].

The broad substantive and extraterritorial scope and the depth of detail make the Draft an extremely important document on a global scale, with the potential to have a major impact on the regulation across many countries [Greenleaf G., 2021: 9]. This trend has previously characterised other acts of the European Union and has been referred to in the academia as the Brussels Effect[6] [Balford A., 2012: 19].

It is also worth noting that technology companies planning to place their AI products on the EU market are looking for a policy on the development

[3] See ideas on different groups of legal requirements for AI systems depending on the potential risk of their application, which are contained in the most important documents characterising the US approach: Blueprint for an AI Bill of Rights. Available at: URL: https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (accessed: 08.10.2023) и NIST AI Risk Management Framework. Available at: URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (accessed: 08.10.2023). The need for a risk-based approach has also been repeatedly expressed at US Congressional hearings on new legislative initiatives. The same approach is also reflected in the bill introduced in September by Senators R. Blumenthal and J. Hawley's Bipartisan Framework for U.S. AI Act. Available at: URL: https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf (accessed: 08.10.2023)

[4] See Artificial Intelligent White Paper 2022 describing China's regulatory approach and including a provision combining a risk-based approach with the level of autonomy (the proposal is to establish three groups of AI systems according to their level of autonomy and three groups according to the risk of their use in relation to human rights). Available at: URL: https://cset.georgetown.edu/publication/artificial-intelligence-white-paper-2022/ (accessed: 08.10.2023)

[5] The Concept for the Development of Regulation in Artificial Intelligence and Robotics Technologies until 2024 explicitly states that it is premised on a risk-based and human-centred approach. The Code of Ethics for Artificial Intelligence contains similar provisions. Available at: URL: http://government.ru/docs/all/129505/ (accessed: 08.10.2023)

[6] The Brussels effect refers to the unilateral influence of acts and standards adopted at the EU level on the legal systems of other countries. A similar phenomenon has previously been observed, e.g., in laws on data circulation, antitrust regulation, environmental protection and food safety.

and use of AI systems that will take into account most of the provisions of the Draft to facilitate future compliance. Moreover, developers are already partly taking these requirements into account. For example, a recent study by a group of scholars from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) evaluated, using twelve criteria, how well the most advanced foundation models currently meet the requirements of the Draft. The authors of the study concluded that the degree of compliance with the act varies widely from 25% to 75%. However, meeting all or most of the legal requirements is quite feasible, which will help to improve the quality of functioning and product safety[7].

Thus, in view of the fact that the AIA Draft is the most comprehensive initiative to date, a study of its approaches is essential for balanced regulation, including regulation in the Russian Federation, because Russia, like most other states, has not yet moved from the stage of approving concepts to the development and adoption of laws and regulations. At the same time, the key principles underlying the Concept for the Development of Regulation in Artificial Intelligence and Robotics Technologies until 2024 approved by the Decree of the Government of Russia[8], are, for the most part, similar to those contained in the Draft mentioned. Also, technology businesses planning to participate in the international market in the future should understand the development of global regulatory trends.

The authors of the paper aimed to explore the risk-based approach contained in the Draft, identify the main regulatory legal requirements imposed on entities placing AI systems on the market, and analyse the key challenges facing the legislator at this stage. The results of this research can be used by government agencies in the development of concepts and regulations, as well as by businesses in preparing to meet the requirements for placing AI systems on the markets.

A series of general and specific scholarly methods were applied in the course of the work. The analysis method was used to divide the Draft and other statutory acts into separate parts, which allowed for a detailed examination of their structure and internal elements. The synthesis method was used to combine the internal elements of the reviewed documents into single semantic blocks, which contributed to obtaining comprehensive knowledge

[7] See: Do Foundation Model Providers Comply with the Draft EU AI Act? Available at: URL: https://crfm.stanford.edu/2023/06/15/eu-ai-act.html (accessed: 08.10.2023)

[8] Available at: URL: http://government.ru/docs/all/129505/ (accessed: 08.10.2023)

about the subject matter under study. The induction and deduction methods helped to identify common features and differences characteristic of the way the risk-based approach is applied in various countries and regions. The systematic approach helped to systematise and structure the knowledge about the subject matter under study. The formal legal method was used to study the provisions of individual legislative acts, which helped to determine the features of legal regulation of public relations in the area under consideration. The comparative legal method was used to identify the advantages and disadvantages of the risk-based approach stipulated in the Draft.

The Risk-Based Approach in the Draft: Features and Key Challenges

1. The Concept of AI Systems and Their Classification by Risk Levels

1.1. AI Systems Definition in the Draft

To begin consideration of the AI systems and the way they are classified by the risk level, we have studied their definition given in the Draft. This is essential for understanding what particular products potentially fall within its scope. The latest version of the document[9] offers the following definition: "Artificial intelligence system (AI system) means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments."[10]

[9] All the three versions of the Draft contain definitions of the term 'AI system' that slightly differ from each other. The European Commission text (2021): "...software that is developed with one or more of the techniques and approaches (these approaches are listed separately in an annex to the document) and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with." The EU text: "...a system that is designed to operate with elements of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve a given set of objectives using machine learning and/or logic- and knowledge based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts."

[10] Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Available at: URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (accessed: 30.08.2023)

As this definition is quite broad, it allows including into AI systems a large number of software products developed on the basis of various methods and techniques, and not only those based on neural networks or machine learning techniques. Technology neutrality is another important feature. AI systems are defined through essential attributes that are inherent to them rather than by listing relevant technologies and methods. It should also be noted that the definition under review was an intentional move by European legislators towards terminology unification at the international level. For example, the Recommendations of the Council on Artificial Intelligence of the Organization for Economic Co-operation and Development (OECD) contain a similar definition.[11] Currently, this version is the most widespread and has become the basis for regulatory concepts in many OECD countries (including such leaders in the field of AI technologies as the USA[12]).

The approach to AI systems definition that aims to identify their main attributes is the most flexible of all and is justified for a legislative document. The attributes in question include: tasks performed, human role in tasking, operating environment, autonomy, and self-learning. More concrete recommendations on AI systems classification that are not technologically neutral may be in the future included in technical standards and in enactments issued by executive authorities [Schuett J., 2023: 3].

At the same time there is a variety of AI systems that can be used in completely different scenarios, from recommendation generation and content creation to critical infrastructure management and national security. Consequently, a specific set of means and methods of legal impact should be applied to different groups of such systems.

1.2. Classification of AI Systems by Risk Levels

The Draft under review uses a risk-based approach to classify AI systems into groups. The higher the risk of human rights violations from the

[11] Note: the OECD document contains the following definition: "An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy." AI system: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy. Available at: URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed: 30.08.2023)

[12] NIST AI Risk Management Framework (AI RMF 1.0). Available at: URL: https://www.nist.gov/itl/ai-risk-management-framework (accessed: 30.08.2023)

use of individual AI systems, the more stringent the requirements placed on them. The Draft provides for a total of four such groups: prohibited AI systems, high-risk AI ones, limited-risk AI ones, and low-risk AI ones. Each group of AI systems has its individual legal requirements.

The Draft applies to entities operating AI systems in the EU. 'Providers' who deploy such systems in the EU market are among such entities, and it does not matter where they are domiciled or actually located. The decisive factor is whether the results of these system operation are intended for use within the EU. Even if the provider is in a third country but uses output data in the EU, it will fall under provisions of the Draft. The document then uses the term "deployer" of an AI system; however, what it means is not the end user but entities using an AI system at other levels (downstream usage). This is supported by the provision that deployers are individuals who do not use such systems for personal (non-professional) purposes. In addition, the original version of the document used the term 'user', and the current version uses the term 'deployer.' In this way, lawmakers sought to stress that they meant specifically entities using AI systems in their products. The Draft also applies to importers, distributors, authorised representatives of providers and manufacturers of products. Such entities — unlike providers and users — must be located or registered in the EU.

One disadvantage of the risk-based approach is its inflexibility: as technology evolves, the classification of AI systems will have to be revised frequently.[13] Experts suggest that this problem could be somewhat mitigated, in particular, by using a more flexible approach to categorising AI systems into groups based on the risk. Their risk assessment system consists of two steps: the development of risk scenarios and the application of a proportionality test. Such an approach may improve the application of the Draft AIA [Novelli C. et. al., 2023: 4–5].

At the same time, it is possible that dividing regulatory requirements for AI systems only into risk groups will not address all of the challenges facing lawmakers. For example, the text of the Draft proposed by the European Commission did not allow regulating the market entry of foundation models[14] and general purpose AI systems[15] whose wide-scale use began in a large

[13] The regulation of artificial intelligence. Available at: URL: https://link.springer.com/article/10.1007/s00146-023-01650-z (accessed: 08.10.2023)

[14] The Draft gives the following definition for the foundation model: 'foundation model' means an Al system model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks.

[15] The Draft AIA gives the following definition for the general purpose AI system:

number of fields only at the end of 2022. A big part of the problem is that the Draft focuses on establishing responsibilities for various entities that are going to place systems on the market. At the same time, other distribution channels are usually typical of foundation models. For example, the most powerful and popular channels are privately owned. Companies provide access to their use and customisation for commercial purposes for a fee through software interfaces (APIs). That means some companies build and deploy these systems, while others apply them to solve a wide range of tasks. However, the latter group do not have access to the full source code of the model, the training data, or the infrastructure (sometimes this can be third-party cloud computing power); nor can they improve or adjust the model. Hence, it is not possible to use an approach that focuses all attention only on the actors that actually place AI systems on the market. Thus, it is necessary to establish regulatory requirements for all stages of the life cycle of AI systems, such as development, deployment, and application.

The present level of foundation models opens up a broad range of opportunities for the creation of autonomous agents on their basis in the coming years, and such agents would be capable of undertaking individual activities, including legally significant ones, on behalf of a human. So, it has a sense to look at the level of AI system autonomy as one of the areas that requires legal regulation.

Thus, to work out a more flexible regulatory approach, we need to differentiate requirements both by risk levels and by all stages of the life cycle of AI systems and the degree of their autonomy. This classification will make AI systems more flexible, that will allow to apply a wider range of legislative requirements. For example, it will become mandatory to test some systems mentioned and foundation models in regulatory sandboxes before placing them on market; some such systems will have to undergo external independent audits; others will have to undergo internal compliance assessments. Additionally, the law may require that to place some AI systems on the market, internal ethical and corporate standards and risk management frameworks must be established.

2. Regulatory Requirements for Certain Groups of AI Systems

2.1. General Principles Applicable to All AI Systems

'general purpose AI system' means an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed.

The Draft establishes a list of general principles to guide operators (providers etc.) at all stages of development and operation of AI systems and foundation models. These principles include: human agency and oversight; technical robustness and safety; privacy and data governance; transparency, diversity, non-discrimination and fairness; social and environmental well-being. These guidelines do not, however, directly impose additional legal obligations on operators. The meeting of the Draft specific requirements that relate to different AI system types and foundation models will for the most part serve as evidence of their compliance. These principles should then be incorporated into technical and corporate standards. Moreover, the Draft explicitly places the obligation to include them in technical standards on the European Commission and the future AI Office[16]. These documents will help to develop rather abstract principles into technical requirements.

2.2. Prohibited AI Systems

The risk-based approach stipulates a separate group of AI systems that, by virtue of their functional characteristics, pose an unacceptable risk to human rights and freedoms. For this reason, their use is illegal in the EU. The Draft identifies several groups of prohibited uses of AI systems.

It is prohibited to use these systems that (in a covert manner) manipulate a person's behaviour so that this results in material harm to her/him or another person. This prohibition will apply to AI systems, which simultaneously meet the following criteria: the system influences the person in question at the subliminal level or performs deliberate manipulation; the person makes an uninformed decision; the system causes substantial harm. The initial version of the Draft stipulated that this prohibition applies to all cases where physical or psychological harm is caused. This understanding was too narrow because AI systems can also cause social, cultural, financial and other harm. [Neuwirth R., 2023: 6–7].

The Draft also prohibits AI systems to make use of vulnerable human attributes (age, disability, etc.) resulting in behavioural change and substantial harm. In other words, it is illegal to use AI systems to classify individuals by using legally protected sensitive attributes.

Social scoring of individuals (groups of individuals) is placed in an independent group of prohibited practices. It is not permitted to assess a

[16] A new European Union body to be established under the current text of the Draft. The document defines its intended competence and structure.

person on the basis of their social behaviour or known or predicted personality characteristics. Such an assessment must result in discriminatory treatment of certain individuals (groups): (a) in a social context unrelated to the context in which data about them were originally generated or accumulated; or (b) that is unjustified or disproportionate to their social behaviour or its severity.

The list of prohibited scenarios for the use of AI systems also includes: use of remote real-time biometric identification systems in public places; use of predictive analytics to determine the likelihood of an individual committing an offence; creation of databases based on untargeted collection of facial images from the Internet or CCTV footage; use of emotion recognition software in law enforcement, border control, educational institutions, and at the place of work.

And, finally, video footage from publicly accessible locations may not be analysed using remote biometric identification systems unless such use is subject to judicial authorisation under EU law for the purposes of a search (of persons) related to a criminal offence.

From the point of view of applying above prohibitions, the provisions that do not allow the use of subliminal influence techniques are a challenge [Neuwirth R., 2023: 3]. It is clear that subliminal techniques can significantly influence decision making and lead to undesirable consequences for the individual. At the same time, the term "subliminal" is difficult to define, and the Draft gives no explanation of its meaning. AI systems can often influence human behaviour using both conscious and subliminal techniques at the same time. For example, smart glasses can influence the human psyche in an overt way by showing pictures, videos, playing music, and, at the same time, in a covert way, read the person's emotions through eye movement recognition, electrical activity in the brain, heartbeat and heart rhythms, muscle activity, blood density in the brain, blood pressure, and skin conductivity.

As a result, it would be difficult to establish whether subliminal techniques have been used, and that these techniques have caused a significant distortion of a person's behaviour. The Draft or other acts should clearly define the term "subliminal techniques" and clarify the legality of their use.

The issue of classifying certain systems as prohibited has been a matter of debate among political forces due to the difficulty in balancing human rights and the public interest. Not everyone who participated in the discussions

were satisfied with the results of the consensus reached after the text was approved by the European Parliament. In particular, human rights organisations asked the EU bodies to be more diligent in protecting human rights during the trilogue. For example, one of the proposals was to involve civil society actors in assessing the impact of AI systems on fundamental human rights, provide for the possibility to appeal decisions taken by AI systems, including through human rights defenders, and establish flexible compensation for victims. It was also proposed to introduce restrictions on the use of AI systems in law enforcement, migration control, and national security.

Thus, legislators should formulate clear criteria for classifying AI systems as prohibited. It will allow developers to better understand the permissible boundaries when creating products, on the one hand, and avoid arbitrary classification of systems as prohibited by law enforcement authorities, on the other.

2.3. High-Risk AI Systems

Title III of the Draft lists the requirements to high-risk AI systems. According to Article 6, AI systems listed in Annexes II and III belong high-risk AI systems, independently or as a component of the safety system of another product.

Annex II contains two lists of acts of the harmonised EU laws, those based on the New Legislative Framework, and others. The acts categorised under this Annex define products and areas of the economy in which the application of AI systems is associated with increased risk. Annex III sets out eight groups that categorise AI systems as high-risk systems by the areas of their application. These include, among others: biometric and biometrics-based systems; AI systems for the management and operation of critical digital infrastructure; AI systems for education and vocational training. Together, these Annexes are intended to provide an exhaustive list of high-risk AI systems by allowing for the inclusion of large areas of the economy as well as more specific usage scenarios.

The EC will develop updated requirements for categorising such systems after consultation with the AI Office at least six months before the Draft enters into force. Law-enforcement agencies in the EU have enough time to make final and balanced decisions so as not to impose excessive requirements and in this manner stifle entrepreneurial activity.

There is a new layer of regulation in the current version of the document that significantly reduces the list of systems, which can be categorised among high-risk systems. For instance, high-risk AI systems identified on the basis of the areas of their application (Annex III) will now only be recognised as such if they significantly threaten life, safety and fundamental human rights. AI systems for managing and operating critical digital infrastructure must additionally pose a significant risk of harm to the environment. Introducing this layer of requirements was a major step towards liberalising business requirements. This has significantly reduced the list of AI systems that will be classified as high-risk systems.

The Draft stipulates a number of requirements that must be met for high-risk AI systems to be placed on the market. A risk management system must be established and implemented, and then needs to be updated in a timely manner throughout the life cycle of the AI system; data sets (training, validation and testing data sets) for the AI systems that are based on such systems should be quality tested; all necessary documents about the system must be created and updated in a timely manner before the system is placed on the market; the system should be able to record all activities during its operation in a special logbook; the operation of the system should, as far as possible, be understandable and transparent to different levels of providers and end users; systems should be designed to be controllable by a human being; systems should be designed from the outset to meet the requirements of safety, reliability, accuracy, resilience and cybersecurity. Alongside the above provisions, additional requirements are placed on individual high-risk AI systems. For example, these must be registered in a single database and must undergo the fundamental rights impact assessment for high-risk AI systems.

Conformity assessment, as envisaged in the Draft, is an integral part of high-risk AI systems' safety and reliability. Providers of high-risk systems must undergo this procedure before releasing their product to the market. There are two types of conformity assessment procedures: (a) the conformity assessment procedure based on internal control referred to in Annex VI; (b) the conformity assessment procedure based on assessment of the quality management system and assessment of the technical documentation, with the involvement of a notified body[17], referred to in Annex VII.

[17] Notified body means a conformity assessment body notified in accordance with the Draft and other relevant EU harmonisation legislation.

This second type of procedure will be used in a relatively limited number of cases where either technical standards and common specifications developed by the European Commission are not applicable, or the supplier voluntarily decides to undergo an external conformity assessment regardless of the categorisation of the AI system under a particular risk level. A voluntary conformity assessment by a notified body can be a competitive advantage, as it will mean that the public agency has guaranteed product safety to consumers. Such an incentive will help improving the overall quality of AI systems without introducing additional stringent regulatory measures.

The Draft has been repeatedly criticised, and it has become the subject of scholarly discussions in the context of conformity assessment procedure. In particular, a group of experts noted that the Draft did not provide detailed explanations on how such an assessment should be undertaken [Mökander J. et al., 2022: 251]. The guidelines developed to date in academia can significantly help businesses overcome this shortcoming. Examples of such documents include: capAI — a guide to going through this procedure, which documents in detail all the measures that high-risk AI system providers need to take[18]: Guidelines for assessing the ethics and reliability of AI systems at different stages of their life cycle in determining the intended use, design, and development [Vetter D. et al., 2023: 5].

Another point of debate is that effective verification of AI systems requires an external independent audit based on ethical principles and standards [Mökander J. et al., 2021: 21–22]. Scholars note that not only lawyers, engineers and philosophers, but also specialists in the field of management should be involved in the development of audit procedures. This conclusion was based on the experience of auditing AstraZeneca's AI systems for ethical compliance. The authors of the study showed that the main difficulties organisations face in auditing AI systems are related to usual management problems. They also touched upon questions of the audit structure. For instance, the authors proposed a 'three-layer' audit for large language models: audit of management, audit of the model, and audit of its application [Mökander J. et al., 2023: 5, 464].

Thus, classifying a small group of systems as high-risk AI systems is a positive measure aimed at creating favourable conditions for business and innovation development. The same applies to the conformity assessment procedure,

[18] CapAI — A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. Available at: URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091 (08.10.2023)

which in the vast majority of cases will be conducted on the basis of internal control. It seems, however, that some of the most powerful AI systems and foundation models may eventually require more stringent requirements, such as external independent auditing and licensing, to place on the market.

2.4. Limited-Risk AI Systems and Low-Risk AI Systems

This group of AI systems should meet additional requirements for operational transparency (Title IV). For example, providers should ensure that all necessary measures are in place to make it clear to users that they are interacting with AI systems. They should also provide information on the permissible functions of the AI system, human control over it, the entity making the final decisions, and the procedures for challenging these decisions in accordance with the law. Providers of authorised systems that recognise human emotion should seek consent to process biometric information of the individuals in question. It is also stipulated that 'deepfakes' must be labelled — unless the content is obviously generated for artistic, humorous or other purposes.

The main idea behind these provisions is that individuals should be informed about their interactions with AI systems. For example, they need to know that their emotions or other characteristics are being recognised, or that image, video or audio content is being generated. This will increase public confidence in AI systems [Chamberlain J., 2023: 5].

Title IX of the Draft stipulates that developers of such AI systems are encouraged to elaborate voluntary Codes of Conduct that reflect how the principles envisaged for all the AI systems discussed earlier are to be implemented. Then it will be clear to users how to operate the system correctly and what measures the developers have taken to make the products safe.

At the same time, researches show that the perception of AI systems and the effect of their application depends very much on what information the user has about them [Pataranutaporn P. et al., 2023: 3]. It is quite easy to mislead people and lower their alertness through proper advertising and overly positive product descriptions. Thus, there is a need to demand that companies develop adequate and understandable rules for the use of the AI system that contain notifications of possible negative consequences. The same should apply to the interfaces that users interact with.

3. The Risk-Based Approach in the Context of Foundation Models and General Purpose AI Systems

The key issue in the finalisation of the Draft is the choice of regulatory approach to the development and application of foundation models and general purpose AI systems. The three versions of its text contain different provisions: only general requirements that apply to all AI systems by risk level and no additional requirements (European Commission text); additional requirements are established for general purpose AI systems (European Council text); individual requirements are established for foundation models, while general purpose AI systems are subject to general requirements on risk levels (European Parliament text). All of the approaches have a number of debatable and ambiguous controversial provisions. Considering the high relevance of the content of the Draft for political forces, business, and the public, it is still difficult to predict unequivocally whether any of the approaches considered will be chosen as the main one, or whether the final text will to some extent combine all of them. Moreover, in some cases, finding the most balanced solution is complicated by the lobbying of large technology companies[19] that have the power to influence the process of drafting and discussing regulations.

Scholars have also taken other positions on the place of general purpose AI systems and foundation models in a risk-based approach. For instance, researchers at The Future Society[20] suggest that all general purpose AI systems should be categorised into three broad groups based on the levels of risk they pose to human rights: Generative AI systems (400+ providers); Group 1 general purpose AI systems (foundation models) (~14 providers); Group 2 general purpose AI systems (frontier foundation models) (~10 providers). Each group will have a different set of legal requirements. Group 3 will be characterised by the most extensive regulatory requirements, which include, in addition to the requirements for all other groups, requirements such as: internal and external independent audits, regular interaction with the AI Office, full transparency, etc. At the same time, this approach is clearly weak as it offers a division into too few groups and is too reliant on current technological realities.

Thus, we believe it is necessary to divide the requirements for foundation models and general purpose AI systems into different groups. Specific requirements should be applied to foundation models, taking into account

[19] See: The lobbying ghost in the machine. Big Tech's covert defanging of Europe's AI Act. Available at: URL: https://corporateeurope.org/sites/default/files/2023-03/The%20Lobbying%20Ghost%20in%20the%20Machine.pdf (accessed: 08.10.2023)

[20] Heavy is the Head that Wears the Crown. Available at: URL: https://thefuturesociety.org/heavy-is-the-head-that-wears-the-crown/ (accessed: 08.10.2023)

that different actors will distribute them at all stages of their life cycle. Requirements for general purpose AI systems should vary based on a risk-based approach. Placing the frontier general purpose AI systems on the market must be based on more extensive regulatory requirements.

The latest version of the Draft retains the term "general purpose AI systems", but extends the requirements to the development and application of the foundation models and the entire AI value chain. The new Article 28b established a number of requirements for AI systems that they must meet before they can enter the market. These include: take measures to mitigate possible negative consequences from their application, use pre-trained and validated data sets, develop only models that can be safe, transparent and predictable throughout their lifecycle, keep relevant technical documents about the model for at least 10 years from the date of its release to the market, etc. Generative AI systems must meet additional requirements: comply with transparency requirements, build and train models in such a way that they cannot potentially be used for infringing purposes, and disclose details of the use of copyrighted material in datasets. All these measures are designed to place additional obligations on the developers of AI systems and thereby offset the shortcomings of the risk-based approach that involves only setting requirements for entities bringing AI systems to market.

The requirement to disclose datasets causes the greatest controversies. This issue is extremely painful because its regulation requires a balance between support for content creators and technology development [Hacker P., 2021: 259]. At the same time, the Draft stipulated long lead times for the preparation of datasets by technology companies when these create new products. Some companies already voluntarily use only legally clean data to create their products nowadays[21].

Another important measure that is widely discussed in academia and society is the right of an individual to prohibit the use of their data or their property to train AI systems. Requirements in this regard have not yet been reflected in the Draft, but some people in the business community have expressed their willingness to offer such waivers[22].

[21] Adobe's Firefly has been fully trained on legally clean data (on its Adobe Stock dataset and on open licence works and public domain content whose copyrights have expired). Also, the company has a whole team of moderators who check new data for copyright infringement risks before adding it to datasets.

[22] For example, StabilityAI voluntarily accepts applications from authors demanding that their content be removed from datasets. OpenAI has announced that it will not collect

Another pressing issue is access to AI systems and foundation models. The current text only allows to test them in regulatory sandboxes. Meanwhile, legislating an obligation to leave open access to AI systems and foundation models for scholars and researchers would be a rational measure. This would ensure the necessary level of transparency in the functioning of such systems because independent experts could monitor the quality of AI systems and identify potential threats in a timely manner.

A number of issues regarding the distribution of foundation models and AI systems under open licences also remain unspecified. In particular, a group of companies that distribute advisory software have suggested that lawmakers should provide a clear definition of AI components. The latest version of the text of the Draft (European Parliament version) contains such a term regarding open-source (Articles 5e и 12a-c), but does not give it an exhaustive definition.[23] Another rational solution in helping small businesses may be to differentiate requirements for foundation models suppliers depending on their use cases, development methods, and market position. Scholars suggest using, e.g., a staggered system for bringing foundation models to market. It implies that hazard levels of the system should be defined to grant access to the system under open licences [Solaiman I., 2023: 119]. This means that, e.g., foundation models with market-leading features will be prohibited for distribution via open-source due to high risks of leakage and misuse.[24]

Conclusion

Although the Draft has been actively developed and discussed for several years, there are still a number of issues that have not been clearly resolved. Moreover, the constant changes in technology create new problems

---

data labelled "Do Not Train". A whole range of US companies that are part of the Content Authenticity Initiative have developed and are implementing Content Credentials. The technology allows for the addition of a "Do Not Train" tag to metadata, which should allow the data not to be included in future datasets, digitally tag the data for authorship, and separate generated content from copyrighted content (in order to protect human-created elements with copyright).

[23] Supporting Open Source and Open Science in the EU AI Act. Available at: URL: https://huggingface.co/blog/assets/eu_ai_act_oss/supporting_OS_in_the_AIAct.pdf (accessed: 08.10.2023)

[24] Open-Sourcing Highly Capable Foundation Models. Available at: URL: https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models (accessed: 08.10.2023)

and challenges for lawmakers. In addition, the extensive adoption of foundation models over the past year requires refinement of the original concept of the risk-based approach.

In order to build a flexible regulatory approach, requirements need to be differentiated both by risk levels and across all stages of the life cycle of AI systems and their degree of autonomy. This will allow a wider range of legislative requirements to apply to different groups of systems. This approach also makes it possible to take into account the distribution of these systems and foundation models by different actors and to properly regulate all stages of their life cycle.

The provisions related to the classification of such systems by risk levels need to be refined. First, the range of prohibited systems should be clearly defined on the basis of clear criteria. It will help developers to better understand the regulatory requirements, and to eliminate arbitrary practices in the decisions taken by law enforcement agencies. Second, classifying a small group of systems as high-risk systems may have a positive impact on innovation and technology development. However, some of the most capable systems and foundation models may eventually need more stringent requirements, such as external independent auditing and licensing, to be placed on the market. Third, legal requirements are needed to develop adequate and understandable rules for the use of systems and their interfaces, which should notify the user of possible negative consequences.

An analysis of the requirements for placing foundation models on the market has shown that the existing approach can be improved by implementing a number of additional regulatory requirements. First, regulatory requirements for foundation models should take into account their distribution by different actors at all stages of their life cycle, and requirements for general purpose AI systems should take into account their risk level. Second, users should be able to unilaterally opt out of having their data used to train these systems. Third, researchers should be given access to the systems and foundation models to ensure their security. Fourth, additional requirements for placing AI systems on the market under open licences should be provided.

References

1. Bradford A. (2012) The Brussels Effect. *Northwestern University Law Review,* vol. 107, no. 1, pp. 1–64.

2. Chamberlain J. (2023) The Risk-Based Approach of the European Union's Proposed Artificial Intelligence Regulation: Some Comments from a Tort Law Perspective. *European Journal of Risk Regulation,* vol. 14, no. 1, pp. 1–13.

3. Gstrein O. (2022) European AI Regulation: Brussels Effect versus Human Dignity? *Zeitschrift für Europarechtliche Studien,* vol. 4, pp. 755–772.

4. Greenleaf G. (2021) The "Brussels Effect" of the EU's "AI Act" on Data Privacy Outside Europe. *Privacy Laws & Business International Report,* issue 171, pp. 3–7.

5. Hacker P. (2021) A legal framework for AI training data—from first principles to the Artificial Intelligence Act. *Law, Innovation and Technology,* vol. 13, no. 2, pp. 257–301.

6. Mahler T. (2021) Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act Proposal. In: Publicerad i Nordic Yearbook of Law and Informatics 2020–2021: Law in the Era of Artificial Intelligence, Mars, pp. 247–270.

7. Mökander J. et al. (2023) Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics,* vol. 3, issue 2, pp. 451–468.

8. Mökander J. et al. (2022) Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds & Machines,* vol. 32, issue 2, pp. 241–268.

9. Mökander J. et al. (2021) Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics,* vol. 27, issue 4, pp. 1–30.

10. Mökander J. et al. (2023) Auditing large language models: a three-layered approach. Available at: https://doi.org/10.1007/s43681-023-00289-2

11. Neuwirth R. (2023) The EU Artificial Intelligence Act: Regulating Subliminal AI Systems. L.: Routledge, 144 p.

12. Neuwirth R. (2023) Prohibited artificial intelligence practices in the proposed EU Artificial Intelligence Act (AIA). *Computer Law & Security Review,* vol. 48, pp. 1–41.

13. Novelli C. et. al. (2023) Taking AI risks seriously: a new assessment model for the AI Act. *AI & Society*, vol. 38, no. 3, pp. 1–5.

14. Pataranutaporn P. et. al. (2023) Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. Nat Mach Intell. Available at: https://doi.org/10.1038/s42256-023-00720-7.

15. Schuett J. (2023) Risk Management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, February, pp. 1–19.

16. Solaiman I. (2023) The Gradient of Generative AI Release: Methods and Considerations. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. N.Y.: Association for Computing Machinery, p. 111–122.

17. Veale M. et. al. (2021) Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, vol. 22, issue 4, pp. 97–112.

18. Vetter D. et. al. (2023) Lessons Learned from Assessing Trustworthy AI in Practice. *Digital Society*, vol. 2, issue 3, pp. 1–25.

Information about the authors:

D.L.Kuteinikov — Candidate of Sciences (Law).
O.A.Izhaev — Candidate of Sciences (Law), Senior Researcher.

Contribution of the authors: the authors contributed equally to this article.

The paper was submitted to editorial office 14.09.2023; approved after reviewing 05.10.2023; accepted for publication 05.10.2023.