УДК 340.132:004.8](4-6ЄС)
DOI https://doi.org/10.32782/pyuv.v3.2024.48

М. В. Шкуратенко
orcid.org/0009-0008-0771-4774
аспірант кафедри цивільно-правових дисциплін
Національної академії внутрішніх справ

СПЕЦІАЛЬНІ ПРИНЦИПИ РЕГЛАМЕНТУ ЄВРОПЕЙСЬКОГО СОЮЗУ З УРЕГУЛЮВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ

Динамічний розвиток штучного інтелекту (далі – ШІ) вимагає відповідного правового регулювання, і якщо, більшість країн все ще перебувають на етапі обговорення, ЄС став флагманом у законодавчій розробці.

12 липня 2024 року Регламент Європейського Союзу з регулювання Штучного Інтелекту (далі Регламент) був опублікований в Офіційному Журналі ЄС, що ознаменувало завершення трирічного легіслативного процесу з його створення [1]. Регламент вступає в законну силу через 20 днів з дати публікації, а отже вже 2 серпня 2024 року майбутнє ШІ та його регулювання, не лише на рівні ЄС, а й у загальносвітовому масштабі, формуватиметься із врахуванням Регламенту. Оскільки Регламент застосовує принцип екстратериторіальності, це означає, що навіть найменший зв'язок системи ШІ ринком ЄС (до прикладу провайдер поза межами ЄС, але система ШІ виходить на ринок ЄС), означатиме застосування до неї тих самих правил, що і для системи ШІ в межах ринку ЄС (провайдер нестиме відповідальність за відповідність системи ШІ до законодавства ЄС, якщо остання буде присутня на ринку ЄС, але місцезнаходження самого провайдера поза межами ЄС). Регламент законодавчо закріплює визначення «система ШІ», яке є певною мірою загальноприйнятим, і гармонує з Виконавчим ордером щодо надійного, безпечного та довірливого розвитку та використання ШІ, Президента Байдена [2] та AI Papers Організації економічного співробітництва та розвитку [3]. Предмет регулювання Регламенту побудований навколо застосування систем ШІ, які можуть становити найвищий ризик. «Ризик» Регламент визначає як комбінацію можливого настання шкоди та вагомість шкоди. Такий підхід дозволяє в певній мірі відповідати головному виклику, пов'язаному з регулюванням ШІ: законодавчий процес не встигає та потенційно не зможе встигнути за динамічним розвитком ШІ, а отже варто створити сталий механізм регулювання, який буде актуальним незалежно від змін та підтримуватиме головні принципи Регламенту.

З початкових етапів розробки Регламенту, Пропозиція Єврокомісії щодо регулювання ШІ [4] була побудована на бінарному підході, щодо визначення ризиків, які можуть становити системи ШІ. Таким чином пропонувалося класифікувати системи ШІ виключно як такі з високим та низьким ризиком. В процесі розробки та обговорення, відбувся перехід до чотирьохступеневого підходу, який складає основу прийнятого Регламенту: заборонені системи ШІ (неприйнятний ризик), системи ШІ, які становлять високий ризик, системи ШІ, які становлять обмежений ризик (ризик транспарентності) та системи ШІ мінімального ризику. Чим вищий ризик, тим більше обов'язків покладатиметься на провайдера (деплоєра, імпортера) під час виходу системи ШІ на внутрішній ринок ЄС. Саме класифікація ризиків на чотири категорії має потенціал створити баланс між пропорційністю регулювання систем ШІ та ефективністю, якої має на меті досягнути Регламент. Враховуючи те, що Регламент виступає лише законодавчою канвою, яка не покриває і не має на меті покрити всі види ШІ, до предмету регулювання останнього не входять моделі відкритого програмного забезпечення (далі – open sourse), якщо вони не застосовуються в контексті високого ризику. Прикладами open sourse моделей є Stable Diffusion, Meta LLAma 3, Mistral AI. З одного боку, абсолютно логічним є виключення великих мовних моделей (далі LLM), які є безкоштовними та доступними для всіх з предметного регулювання Регламенту, а з іншого – це має потенціал створення прогалини в законодавстві, якщо, до прикладу, відбуватиметься злиття open sourse LLM, з гігантами машинного навчання, які не є open sourse.

Концепт «надійного штучного інтелекту» (trustworthy AI, той, якому можна довіряти) покладений в основу Регламенту ШІ та безпосередньо корелює з прийнятними та неприйнятними ризиками, які категоризують системи ШІ, а отже і визначають межі їхнього регулювання згідно Регламенту. Довіра є важливою рисою гарно функціонуючого та процвітаючого суспільства, тому не дивно, що одна з регулятивних цілей Регламенту та законодавства в цілому – це створення довіри за допомогою правового регулювання. Ніколас Луманн, один з перших системних дослідників поняття «довіра» в соціумі, визначав останнє як «механізм для навігації

комплексності, який дозволяє дію поза межами миттєвої впевненості» [5]. Френсіс Фукуяма визначає довіру як «соціальний капітал», необхідний для економічного успіху країни. Деніел Халт співвідносить довіру та надійність в праві, як: надійність (trustworthiness) не несе в собі такого ризику як довіра, бути надійним (тим, кому довіряють) означає компетенцію та доброчесність виконати певну дію, і сама наявність цих рис не є ризиком [6].

В 2020 році, Єврокомісія видала White Paper щодо ШІ – Європейський підхід до досконалості та довіри, де закріплюється прагнення ЄС до підтримки розвитку технологій та побудови довіри як основи для швидкого поширення ШІ, враховуючи основоположні цінності та принципи ЄС, а головне сприяння довірі до ШІ. Юридична доктрина не має єдиної відповіді на те чи може ШІ бути справжнім/щирим об'єктом довіри: чи може фізична особа довіряти алгоритмічній системі? Генрік Скауг Сетра звертає увагу на визначення поняття «довіра» у інтеракції «людина-комп'ютер» та його відмінність від базового поняття інтеракції «людина-людина». Головним критерієм у такій взаємодії є наявність характеру (моральної природи, ethos) учасників, тобто для побудови довіри машини теж мають ethos (не відбувається посилання на суб'єктність ШІ) [7]. Надійний ШІ пріоритизує безпеку та транспарентність для тих, хто з ним взаємодіє, а отже, враховуючи те, що немає ідеальної моделі ШІ, яка б не припускалася помилок та була повністю без упереджень, дотримання принципів надійного ШІ є необхідністю, яка сприятиме побудові довіри та обґрунтовуватиме ризик – заснований підхід. Разом із тим, не варто забувати про овертраст, або довіру поза межами необхідного, яка виникає, коли людина починає занадто довіряти алгоритмічній системі, і не контролює output, що призводить до можливих помилок та упереджень [8]. Саме з таких причин Регламент вимагає супервізію фізичної особи над певними системами ШІ.

Людиноцентризм, забезпечення фундаментальних прав та свобод людини, демократія, захист навколишнього середовища, сприяння інноваціям та захист персональних даних, – важливі загальні принципи, закріплені у Європейській хартії основоположних прав, які також складають основу Регламенту. Але разом із тим в Регламенті є і спеціальні принципи, які мають на меті підвищити рівень довіри до ШІ: транспарентність, людський контроль, технічна точність та безпека, тощо. Ці принципи є вагомою складовою побудови надійного ШІ.

Транспарентність як критерій застосовується для систем високого ризику згідно статтей 17, 20 та 21 AI Act, які передбачають розробку та побудову систем ШІ таким чином, щоб їхні операції були достатньо відкритими (транспаретними) та деплоєри системи ШІ могли інтерпретувати output, використовуючи його належним чином. Саме транспарентність модулює основу довіри до ШІ. В 2019 році Експертна Група Європейської Комісії Високого Рівня випустила етичні рекомендації щодо побудови надійного ШІ (AI HLEG), де були визначені сім основних критеріїв надійного ШІ, серед яких транспарентність, яка включає в себе можливість відслідковувати, пояснювати та комунікувати [9]. Вимогу до відкритості також часто пов'язують з принципами пояснюваності (explicability principle), зрозумілості (intelligibility), ясності та інтерпретованості, адже всі вони охоплюють дещо нове про ШІ: робота ШІ є часто невидимою та незрозумілою для більшості (окрім невеликої кількості експертних спостерігачів) [10]. Саме через пов'язаність із вищенаведеними принципами та невизначений до кінця обсяг розуміння терміну «транспарентність ШІ» відкритість та пояснюваність системи асоціюється із позитивними цінностями, такими як «відкрита інформація», «відкрите джерело», «відкритий код», тощо, але в такої відкритості мають бути свої ліміти, щоб нею не зловживали (наприклад використання інформації з Facebook для політичної агітації) [11].

Обов'язок надавати інформацію та корекційні дії, які застосовуються для систем ШІ високого ризику є частиною принципу транспарентності. У разі невідповідності системи ШІ вимогам Регламенту, корекційні дії провайдерів та повідомлення імпортерів, деплоєрів та уповноважених представників є необхідними. На обґрунтований запит уповноважених органів, провайдери систем ШІ високого ризику повинні надати необхідну документацію, щоб підтвердити відповідність системи ШІ вимогам Регламенту.

Яким чином використовувати повний потенціал ШІ, але при цьому не нанести шкоду суспільству та людям залишається головною дилемою, пов'язаною з регулюванням ШІ. Запропонована в Регламенті супервізія фізичної особи (human oversight) над системою ШІ є одним із варіантів балансу між використанням технології, потенційною шкодою та переважаючою користю ШІ. Позиція щодо нагляду фізичної особи над ШІ є новелою в законодавстві ЄС, яка має на меті ствердження людиноцентризму. Людиноцентричний ШІ не означає, що алгоритм машинного навчання чи система ШІ має думати як людина, чи мати когнітивні здібності як у людини. Навпаки, робити розумні системи людиноцентричними означає будувати системи ШІ для розуміння людських потреб та очікувань, допомагаючи людям розуміти системи ШІ у відповідь [12]. У статті 14 Регламенту, закріплюється принцип супервізії фізичної особи над сис-

темами ШІ високого ризику, а саме: розробка та розвиток систем ШІ з відповідними інтерфейсом інструментів людина-машина, з метою ефективного нагляду фізичною особою у період використання, що дозволить превентувати/мінімізувати ризики, пов'язані зі здоров'ям, безпекою та основоположними правами людини.

Важливою складовою регулювання ШІ є мітигація упередженості алгоритмічних систем, яка залежить не тільки від статистичних та програмних упереджень, але і людських та інституційних. Упередження як критерій не є притаманним виключно ШІ, і наразі неможливо досягнути рівня абсолютної неупередженості [13]. Саме тому в Регламенті у статті 15 від систем ШІ високого ризику вимагається дотримуватися принципів технічної точності та надійності. Також, у разі, якщо система ШІ продовжує навчатися, після виходу на ринок, потрібно превентувати упереджений output від впливу на input майбутніх операцій. Такі системи повинні бути стійкі до помилок та невідповідності, мати бекап план, або план дій у разі не функціонування системи.

Системи ШІ – це комп'ютерні системи, а отже вони успадковують всі ризики кібербезпеки, пов'язані з традиційними діджитальними системами, які функціонують в подібному контексті [14]. Вимога щодо кібербезпеки застосовується до систем ШІ високого ризику у ч.5 статті15 Регламенту, яка має бути забезпечена належними технічними рішеннями та відповідати умовам та ризикам. Такі технічні рішення мають включати міри для превенції, визначення, відповіді, вирішення та контролю над атаками, ціллю яких є маніпулювання інформацією (data poisoning, model poisoning, adversarial examples of model evasion). Принцип кібербезпеки також є частиною системи менеджменту ризиків згідно статті 9 Регламенту, яка передбачає тестування та оцінку систем ШІ високого ризику (з можливим тестуванням в реальних умовах за статтею 60 Регламенту).

Спеціальні принципи, які застосовують в Регламенті створюють основу для надійності ШІ. Разом із тим вплив Регламенту на побудову довіри до ШІ, ми зможемо оцінити виключно після вступу всіх його статей у законну силу. Використання повного потенціалу ШІ є дійсно можливим, але виключно з дотриманням спеціальних принципів: технічної точності, безпеки, кібербезпеки, транспарентності, сепервізії фізичної особи, тощо.

Література

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689 (дата звернення: 30.07.2024).

2. Executive order on the safe, secure and trustworthy development and use of artificial intelligence. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (дата звернення: 30.07.2024).

3. OECD Artificial Intelligence Papers № 9: Explanatory memorandum on the updated OECD definition of an AI system. URL: https://www.oecd-ilibrary.org/science-and-technology/oecd-artificial-intelligence-papers_dee339a8-en (дата звернення: 30.07.2024).

4. Ēuropean Commission White paper on Artificial Intelligence: a European approach to excellence and trust from 19.02.202 COM (2020) 65 final. URL: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en (дата звернення 30.07.2024).

5. Luhmann, N. Vertrauen – Ein Mechanismus der Reduktion sozialer Komplexitдt. Stuttgart: F. Enke Verlag, 1968. 95 с.

6. Hult, D. Creating trust by means of legislation – a conceptual analysis and critical discussion. *The theory and practice of legislation.* 2018. Vol. 6, № 1. С. 1–23. URL: https://www.tandfonline.com/doi/full/10.1080/20508840.2018.1434934#abstract (дата звернення: 30.07.2024).

7. Skaug Sдtra, H. A machine ethos? An inquiry into artificial ethos and trust. *Computers in human behavior.* April 2024. Vol. 153. С. 1–15. URL: https://www.sciencedirect.com/science/article/pii/S0747563223004594#:~:text=Trust%20in%20machines%20partly%20parallels,and%20not%20the%20machine's%20ethics (дата звернення: 30.07.2024).

8. Zerilli, J., Bhatt, U., Weller, A. How transparency modulates trust in artificial intelligence. *Patterns.* April 2022. Vol. 3. С. 1–10. DOI: https://doi.org/10.1016/j.patter.2022.100455

9. HLEG Ethics guidelines for trustworthy AI. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (дата звернення: 30.07.2024).

10. Floridi, L., Cowls, J. A unified framework of five principles for AI in society. *Harvard Data Science Review*. 2019. Vol 1.1. С.1-14. URL: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8 (дата звернення 30.07.2024).

11. Larsson, S., Heintz, F. Transparency in artificial intelligence. *Internet policy review.* May 2020. Vol. 9(2). С. 1–16. URL: https://policyreview.info/pdf/policyreview-2020-2-1469.pdf (дата звернення: 30.07.2024).

12. Riedl, O.M. Human-Centered Artificial Intelligence and Machine Learning. *Cornel University.* 2019. URL: https://arxiv.org/pdf/1901.11184 (дата звернення: 30.07.2024).

13. Schwartz, R., et al. Towards a standard of identifying and managing bias in AI. *National Institute of Standards and Technology*. 2020. Vol.1270, С. 1–86. DOI: https://doi.org/10.6028/NIST.SP.1270.

14. Junklewitz, H., et al. Cybersecurity of artificial intelligence in the AI Act: JRS Science for policy report. Publications Office of the European Commission, Luxembourg, 2023. DOI: https://dx.doi.org/10.2760/271009.

Анотація

Шкуратенко М. В. **Спеціальні принципи Регламенту Європейського Союзу з урегулювання штучного інтелекту.** – Стаття.

У статті висвітлюються спеціальні принципи, які застосовуються в Регламенті Європейського Союзу з регулювання ШІ (далі Регламент): транспарентності, супервізії фізичної особи над системою ШІ, безпеки, технічної точності, та іншим. Метою цієї роботи є проведення аналізу впливу вищезазначених принципів на створення концепту «надійного ШІ» та їх відповідності ризик-заснованому підходу до регулювання алгоритмічних систем. Зважаючи на те, що ідеальної системи ШІ не існує, і алгоритм може бути упередженим та допускати помилки, обґрунтованість включення спеціальних принципів до предметного регулювання Регламенту є беззаперечною. Питання довіри до алгоритмічних систем є дискутованим в науковому середовищі, і відносно точна відповідь на нього з'явиться лише після вступу всіх статей Регламенту в дію та практичного їх застосування. Саме поняття ШІ, якому можна довіряти не є новелою Регламенту – це частина загального бачення майбутнього ШІ та його регулювання згідно з ризик-заснованим підходом, який визначає рівень регулювання згідно Регламенту. Ризик заснований підхід притаманний для регулювання ШІ в США, Канаді, ЄС. Побудова Регламенту, заснована на ризиках, які становить система ШІ, є спробою законодавця відповідати на головний виклик у регулюванні ШІ: динамічності розвитку технологій. Регулювання систем ШІ високого ризику включає більше вимог, і пропорційно, більше спеціальних принципів. Одними із яких є мітигація ризиків пов'язаних із кібербезпекою та технічна точність систем. Важливим положенням Регламенту є екстратерриторіальність його дії. Регламент застосовується до провайдерів, які розміщують системи ШІ на ринку ЄС, незалежно від розташування таких провайдерів в межах чи поза межами ЄС. Отже, загальні поняття Регламенту та його принципи матимуть вплив поза межами країн ЄС. Проведене дослідження підтверджує ефективність застосування спеціальних принципів для створення відкритих та зрозумілих систем, а отже, доступності систем ШІ та їх відповідності основоположному принципу – людиноцентризму.

Ключові слова: штучний інтелект, Регламент ЄС з регулювання ШІ, надійний ШІ, ризик-заснований підхід, спеціальні принципи ШІ.

Summary

Shkuratenko M. V. **Special principles of the EU AI Act.** – Article.

This article considers special principles of the EU AI Act. Such principles are: transparency, supervision of natural person over AI system, safety, technical robustness, etc. The main objective of this article is to analyze the influence of the before mentioned principles on the concept of the «trustworthy AI», and the suitability for the risk-based approach towards regulating AI systems. Emphasizing the fact, that there are no ideal AI system and the algorithm could be biased, or make mistakes, the importance of inclusion of the special principles in the AI Act is immense. Trust towards algorithmic systems is still an ongoing discussion in the scientific community and resolution seems to occur only after the AI Act will enter into force. Then the statistics of practical application will become available. The notion of the trustworthy AI is not a novelty introduced by the AI Act. It is a part of the risk-based approach towards the future of AI. USA, Canada and EU are all opting for a risk-based approach. Such a construct of the AI Act is an attempt of the legislator to address and resolve the main issue: dynamic development of technologies and slower legislative process. Regulation of high-risk AI systems consists of more requirements and proportionally includes more special principles. Some of these are: cybersecurity, safety and robustness of AI systems. One of the main provisions of the AI Act is its extraterritorial application. AI Act is applicable to the providers of AI systems, that are established both within the EU and in the third country, where the AI system is placed on EU market. Thus, the general concepts of the AI Act and its principles will have an impact outside the EU countries. The research conducted states the effectiveness of the special principles application for the creation of intelligibility and transparency in AI systems. This means that the regulation of AI systems supports the main principle of human centrism.

Key words: artificial intelligence, AI Act, trustworthy AI, risk-based approach, special principles of AI.