

УДК 341.1

DOI <https://doi.org/10.24144/2307-3322.2025.92.5.29>

ПРАВО НА ПОЯСНЕННЯ АЛГОРИТМІЧНИХ РІШЕНЬ У СИСТЕМІ ЦИФРОВИХ ПРАВ ЛЮДИНИ

Лабик А.Р.,
*аспірантка 1-го року навчання
спеціальності D9 Міжнародне право,
асистентка кафедри міжнародного права
та порівняльного правознавства
Чернівецького національного університету
імені Юрія Федьковича
ORCID: 0009-0005-1376-6771
e-mail: a.labyk@chnu.edu.ua*

Лабик А.Р. Право на пояснення алгоритмічних рішень у системі цифрових прав людини.

Стаття присвячена дослідженню концепції права людини на пояснення рішень, ухвалених із використанням алгоритмів штучного інтелекту. Активне впровадження автоматизованих систем до таких сфер як освіта, соціальне забезпечення, фінансові послуги, правосуддя та публічне врядування зумовлює значний вплив технологій штучного інтелекту на правовий статус особи. Як наслідок, виникає необхідність формування нових підходів до гарантування і захисту прав людини, які б дозволили запобігти дискримінації, упередженому ставленню та непрозорості під час ухвалення рішень відповідними системами.

У статті досліджено групи помилок, які можуть виникнути під час застосування автоматизованого прийняття рішень судовими органами: технічні, соціальні та помилки, пов'язані з відсутністю чи незрозумілістю пояснень. Проаналізовано передумови зародження концепції права на пояснення алгоритмічних рішень, її зв'язок із теорією «чорної скриньки» штучного інтелекту та зрозумілого штучного інтелекту. Розкрито основні різновиди пояснень, отримання яких гарантує дане право: конкретні та загальні, причинно-наслідкові та нормативні, пояснення *ex ante* і пояснення *ex post*. Проаналізовано підходи до визначення мети права на пояснення - захисту конкретного блага, яким може бути прозорість прийняття рішень, довіра до держави, суб'єктів владних повноважень чи приватних операторів або право на самозахист як таке.

Досліджено європейський підхід до визнання та нормативного закріплення цього права через призму Загального регламенту про захист даних та Закону про штучний інтелект. На основі аналізу положень відповідних актів з'ясовано основні елементи, які системно пов'язані між собою та в сукупності становлять право на пояснення автоматизованих рішень. До таких компонентів віднесено: а) право особи знати про факт використання технологій штучного інтелекту у процесі ухвалення істотного для неї рішення; б) право на змістовне і зрозуміле пояснення винесеного рішення; в) право бути поінформованим про правові наслідки цього рішення; г) право його оскаржити та вимагати втручання людини у відповідний процес.

Ключові слова: міжнародне право, права людини, цифрові права, механізми захисту прав людини, право на пояснення алгоритмічних рішень.

Labyk A.R. The right to an explanation of algorithmic decision-making in the digital human rights system.

This article is devoted to researching the concept of the human right to an explanation of decisions made using artificial intelligence algorithms. The active introduction of automated systems into areas such as education, social security, financial services, justice, and public administration has led to artificial intelligence technologies having a significant impact on the legal status of individuals. As a result, there is a need to develop new approaches to guaranteeing and protecting human rights that would prevent discrimination, bias, and opacity in decision-making by such systems.

The article examines groups of errors that may arise during the application of automated decision-making by judicial authorities: technical, social, and errors related to the absence or incomprehensibility of explanations. It analyzes the prerequisites for the emergence of the concept of the right to explanations of algorithmic decisions, its connection with the theory of the «black box» of artificial intelligence and understandable artificial intelligence. The main types of explanations that guarantee this right are revealed: specific and general, causal and normative, ex ante explanations and ex post explanations. The approaches to defining the purpose of the right to explanation are analyzed: the protection of a specific benefit, which may be transparency in decision-making, trust in the state, public authorities, or private operators, or the right to self-defense as such.

The European approach to recognizing and regulating this right was examined through the prism of the General Data Protection Regulation and the Artificial Intelligence Act. Based on an analysis of the provisions of the relevant acts, the main elements that are systematically interrelated and together constitute the right to an explanation of automated decisions were identified. These components include: a) the right of a person to know about the use of artificial intelligence technologies in the process of making a decision that is significant to them; b) the right to a meaningful and understandable explanation of the decision made; c) the right to be informed about the legal consequences of this decision; d) the right to appeal it and demand human intervention in the relevant process.

Key words: international law, human rights, digital rights, human rights protection mechanisms, the right to an explanation of algorithmic decision-making.

Постановка проблеми. Швидкі темпи цифровізації суспільних відносин та розвиток технологій штучного інтелекту внесли суттєві корективи у звичні процеси прийняття рішень у таких галузях як економіка та державне врядування. Що найважливіше, цифрова трансформація справила значний вплив на існуючі механізми забезпечення і захисту прав людини шляхом спрощення та автоматизації процедур, які ще донедавна вимагали залучення значних людських, фінансових, матеріальних та інших видів ресурсів. Разом з тим, під впливом цифровізації виникли численні ризики, що поставили під загрозу право на захист персональних даних, свободу вираження поглядів, право на рівність, низку процесуальних прав тощо.

Сьогодні питання автоматизованого прийняття рішень має неабияке значення, оскільки відповідні механізми було впроваджено до сфер, які безпосередньо впливають на правовий статус особи, її права, свободи та обов'язки. Зокрема, активне використання алгоритмічних систем спостерігаємо у процедурах надання кредитів фізичним або юридичним особам, соціальних виплат, відбору кандидатів для призначення на різноманітні посади, у конкурсах на вступ до навчальних закладів, отримання окремих пільг чи привілеїв, у роботі органів публічної влади. З огляду на це, цілком очікуваною і справедливою видається вимога про забезпечення прозорості та контролюваності рішень, ухвалення яких відбувається з мінімальною участю людини або взагалі без такої.

У цьому контексті все більшої ваги набуває право на пояснення алгоритмічних рішень як одна із найважливіших гарантій прав людини у цифровому середовищі. Воно спрямоване на подолання так званої «чорної скриньки» штучного інтелекту та створює основу для можливостей особи зрозуміти суть і логіку алгоритмічних процесів, а також отримати доступ до причин прийняття та очікуваних результатів важливих для неї рішень.

Мета дослідження полягає в аналізі правової природи, змісту та визначенні стану міжнародно-правового регулювання права на пояснення алгоритмічних рішень як одного із цифрових прав людини.

Стан опрацювання проблематики. У науковій юридичній літературі дослідженню впливу штучного інтелекту на такі сфери як правоохоронна діяльність, судочинство, державне врядування були присвячені праці В. Гончаренка, О. Явора, А. Тимчишина, В. Хахановського, А. Гачкевича, Н. Петренка, О. Баранова, Л. Машковської, А. Кучук, Ю. Завгородньої та інших. Безпосередньо право на пояснення алгоритмічних рішень досліджували Г. де Брейн, М. Варньє, М. Янссен, Р. Берк, Б. Карлан, Г. Кугельберг, Т. Кім, Б. Рутледж та ін.

Виклад основного матеріалу. Очевидно, операції з автоматизації прийняття рішень мають дещо ризиковий характер, оскільки усунення людини з її власним світоглядом, моральними устоями та ідеями не завжди веде до цілковитої неупередженості. Жодне рішення не може бути ухвалене ізольовано. Причому, на відміну від людини, яку ставлять в рамки її власні знання, використання штучного інтелекту передбачає обробку та аналіз практично нескінченної кількості

джерел, зокрема й тих, які містять суб'єктивну точку зору, упереджені погляди, наслідком чого стають похибки у роботі відповідних систем. Разом з тим, якщо мова йде про бізнес та виробництво, відповідні помилки можуть спричинити фінансові чи матеріальні втрати, проте в питаннях публічного врядування, правосуддя, надання державних послуг тощо, за кожною із таких похибок стоїть доля людини та її права.

Одним із яскравих прикладів ймовірної упередженості систем автоматизованого прийняття рішень можна вважати інструмент оцінки ризиків COMPAS (Система профілювання управління правопорушниками у виправних установах для застосування альтернативних запобіжних заходів), щодо об'єктивності якого точаться суперечки. Це програмне забезпечення, яке активно використовується у багатьох штатах США, було запроваджене з метою полегшення роботи судових органів під час вирішення питань щодо звільнення від відбування покарання, призначення покарання, умовно-дострокового звільнення осіб, які вчинили кримінальні правопорушення. Даний інструмент використовує алгоритми штучного інтелекту для прогнозування ризиків рецидиву на основі ряду факторів, як-то попередні випадки притягнення до кримінальної відповідальності, зловживання алкогольними напоями чи наркотичними речовинами, сімейні та соціальні аспекти. Систему запровадили та активно використовували в Нью-Йорку, Каліфорнії, Вісконсині й інших штатах та округах, поки в 2016 році не була опублікована стаття з обґрунтуванням упередженості відповідних систем в питаннях расової приналежності ув'язнених. Пізніше того ж року в справі «Луміс проти Вісконсина» заявник, вирок щодо якого спирався частково на висновки COMPAS, скаржився до Верховного суду штату на те, що останні були зумовлені його гендерною оцінкою [1, с. 2]. Варто відзначити, що цей інструмент став лише одним із прикладів впровадження технологій штучного інтелекту до правоохоронної діяльності. Разом з тим, багато дослідників скептично ставляться до таких удосконалень, вказуючи на загрозу упередженого ставлення до окремих осіб чи категорій населення. На жаль, часто такі ризики пов'язані з необхідністю саме «передбачення» можливості вчинення правопорушень, за яких аналіз підтверджених фактів становить лише частину від усієї операції, тоді як іншою складовою є прогнозування [2, с. 224].

Україна не стала винятком серед держав, які впроваджують інноваційні технології до систем судочинства та публічного врядування. На сьогодні доступ до значної кількості послуг і процедур був спрощений, а самі операції – удосконалені та переведені в кіберфізичний простір. Разом з тим, юристи-практики виділяють кілька груп проблем, які можуть виникнути у зв'язку з використанням штучного інтелекту при прийнятті рішень правозастосовними органами. По-перше, це так звані технічні або «класичні» помилки, які пов'язані з імплементацією алгоритму, а не його логікою. До них можна віднести неправильне використання мови програмування, помилки в коді, неправильний порядок виконання операцій, обмеженість чи помилковість даних, порогового значення тощо. Друга група – соціальні або правові помилки, які виникають через різного роду упередження, про що вже йшла мова вище. Це найбільш підступний вид помилок, які спричиняють пряму дискримінацію. Вони можуть стати результатом недостатньої представленості окремих соціальних груп, використання даних, які ґрунтуються на суб'єктивній оцінці чи стереотипах, що значно знижує рівень об'єктивності ухвалених рішень і довіри до них громадян. Нарешті, третій тип помилок зумовлений відсутністю або недостатністю аргументації такого алгоритмічного рішення [3, с. 3]. Саме помилки останньої групи дали поштовх для виникнення і розвитку концепції права на пояснення автоматизованих рішень.

На сьогодні ні в науковій літературі, ні в нормативних актах не існує чіткої дефініції права на пояснення алгоритмічних рішень. З огляду на стрімке зростання кількості процесів, що не потребують участі в них людини, та які, як правило, мають істотне значення для особи, дана концепція почала поступово формуватися в різних галузях, набуваючи комплексної природи. Очевидно, ми не можемо обмежити сферу реалізації права на пояснення конкретними рамками, як-то державні послуги, правосуддя чи діяльність правоохоронних органів. Використання технологій штучного інтелекту у роботі та побуті зумовило суттєве розширення змісту даного права, підкреслюючи його фундаментальний та міждисциплінарний характер. Основою для його формування стала концепція «зрозумілого штучного інтелекту» (з англ. - explainable artificial intelligence (XAI)). Вона була розроблена на протидію підходу «чорної скриньки», який повністю ігнорує цінність пояснення алгоритмічного рішення для людини та пропонує покладатися на нього без додаткового обґрунтування. Послідовники теорії XAI, навпаки, вбачають у ньому вищий ступінь розвитку

технологій штучного інтелекту, за якого результати ШІ є доступними та зрозумілими для широкої громадськості, незалежно від складності його моделей [4, с. 1].

Слід відзначити, що правова природа досліджуваної концепції має досить спірний характер. Погоджуючись із тлумаченням права як інструменту для захисту певного блага, дослідники, проте, розходяться в думках, на захист якого саме блага спрямоване право на пояснення. Одні науковці вказують на можливість захистити себе, називаючи її «усвідомленим самозахистом». Згідно з таким підходом, здатність вимагати обґрунтування прийнятого рішення є ключем до ефективної реалізації захисної функції права загалом. Адже чи можна оскаржувати акти органів публічної влади, якщо ми не розуміємо їхніх підстав або ж можливих наслідків? Навряд чи. Ще одна теорія визначає довіру центральним благом, на захист якого спрямоване право на пояснення. Як зазначають її прихильники, завдяки можливості отримати відповідні роз'яснення кожною особою, зростає загальний рівень довіри суспільства до публічної адміністрації. Інші дослідники вбачають такий ключовий елемент і мету визнання права на пояснення у забезпеченні прозорості прийнятих рішень, при тому, що відповідна вимога поширюється як на суб'єктів владних повноважень, так і на представників приватного сектору [5, с. 140].

Розкриваючи комплексний характер права на пояснення автоматизованих рішень, науковці виділяють кілька видів пояснень, на які може розраховувати зацікавлена особа. Так, розрізняють загальні та конкретні пояснення. Як впливає із самих назв, перша група спрямована на тлумачення власне суті того чи іншого алгоритму, механізму ухвалення рішення, його ймовірних результатів. Така інформація має узагальнений характер, проте дозволяє суб'єктові, за наявності достатньої кількості вихідних даних, спрогнозувати, який висновок «видасть» система в його ситуації. На відміну від загальних, конкретні пояснення є індивідуалізованими, роз'яснюючи особі, з яких причин щодо неї було ухвалене відповідне рішення [5, с. 142].

Окрім цього, виділяють також пояснення *ex ante* і пояснення *ex post*, ключовим для розрізнення яких є критерій часу. У той час, як пояснення *ex ante* мають превентивний характер, надаючи необхідну інформацію до початку процесу ухвалення рішення, пояснення *ex post* спрямовані на роз'яснення причин такого рішення уже після факту його прийняття [5, с. 141]. Разом з тим, не можемо погодитись із твердженням про те, що конкретизованими можуть бути виключно пояснення *ex post* [6, с. 80]. Часто такі пояснення можуть бути запізненими, змушуючи особу розпочинати процедуру оскарження чи знову звертатися до уповноважених суб'єктів із тим самим питанням. До прикладу, володіючи інформацією про порядок ухвалення рішення, особа може вагатися щодо одного із пунктів переліку документів, необхідних для цього. Отримавши документ, який не відповідає зразку або є недійсним, система ухвалить негативне рішення щодо такої особи, запобігти якому можна було б за допомогою попередніх роз'яснень, які в даній ситуації мали б конкретний характер. Звісно, конкретні пояснення *ex ante* із об'єктивних причин вміщують меншу кількість індивідуалізованої інформації, проте повністю відкидати можливість їх існування є недоречним.

Разом із наведеними вище, дослідники виокремлюють нормативні і причинно-наслідкові (каузальні) пояснення. Останні, як правило, зосереджуються на факторах (кількісних показниках, інших числових значеннях), які технічно вплинули на остаточне рішення, та розкриваються через контрфактичні пояснення («якби..., то...»). Нормативні характеризуються ширшим змістом і покликані пояснити ухвалене рішення з точки зору його легітимності та обґрунтованості, аж до з'ясування причин застосування систем штучного інтелекту уповноваженим суб'єктом [5, с. 142].

Що стосується регламентації права на пояснення алгоритмічних рішень, як і більшість прав четвертого покоління, та цифрових прав зокрема, дана концепція є недостатньо розробленою. На жаль, визнання та закріплення цього права на рівні законодавств окремих держав є довготривалим та не завжди результативним процесом. Найбільш прогресивним у питанні нормативної регламентації права на пояснення став Європейський Союз. Першим документом Співтовариства, який містив положення щодо зобов'язань контролера надавати необхідну інформацію про рішення, прийняте із використанням автоматизованих систем, став Загальний регламент про захист даних (GDPR). Хоч у GDPR не використовується безпосередньо формулювання «право на пояснення алгоритмічних рішень» або інші його конфігурації, він закріпив архіважливий принцип прозорості в діяльності уповноважених суб'єктів (п. 39), право суб'єкта даних бути поінформованим про використання автоматизованих систем під час прийняття рішень щодо нього, їх логіку, вплив на такі рішення та можливі наслідки (п. «h» частини першої статті 15), право не підлягати рішенню, ухваленому виключно автоматизованими системами, право на його оскарження (ст. 22) [7].

Із набуттям чинності 1 серпня 2024 року Закону про штучний інтелект (Artificial Intelligence Act, AI Act) Європейський Союз зробив ще один крок на шляху до вдосконалення наднаціональних механізмів захисту цифрових прав людини. Стаття 86 цього Закону під назвою «право на пояснення індивідуальних рішень» передбачає можливість особи, щодо якої було прийняте рішення з використанням систем штучного інтелекту із «високим рівнем ризику», вимагати чіткі та змістовні пояснення ролі системи штучного інтелекту в процедурі прийняття рішень та основних елементів прийнятого рішення [8].

Аналіз наведених міжнародних актів дозволяє виділити компоненти права на пояснення алгоритмічних рішень. По-перше, це право особи знати про власне факт ухвалення рішення щодо неї із застосуванням автоматизованих систем. По-друге, складовою досліджуваного права є можливість отримати змістовне роз'яснення щодо суті рішення, передумов, причин, алгоритму його прийняття. Це ключовий елемент права на пояснення, котрий повинен володіти такими характеристиками як доступність, зрозумілість, релевантність, ясність для фахівців та пересічних громадян. Третій компонент - право на втручання та оскарження. Воно закріплює можливість особи оскаржувати рішення, прийняті системами штучного інтелекту, а також вимагати виконання відповідних операцій чи перегляду рішення фахівцем-людиною. Однією із складових права на пояснення є також право бути поінформованим про наслідки ухваленого рішення, тобто з'ясувати, яким саме чином воно вплине на права, свободи чи обов'язки суб'єкта.

Висновки. Як підсумок, варто зазначити, що постійне зростання ролі штучного інтелекту в різних сферах життя суспільства та його вплив на процеси прийняття рішень, які мають юридичне значення, змусили міжнародну спільноту переосмислити традиційні механізми захисту прав людини. Із запровадженням автоматизованих систем, ризики упередженості та дискримінації під час ухвалення рішень нікуди не зникли. Більше того, ці загрози зросли в умовах непрозорості та незрозумілості численних алгоритмів. У цьому контексті право на пояснення становить невід'ємний елемент цифрових прав особи. Воно дозволяє встановити чіткий контроль над автоматизованими процесами не лише з боку конкретних операторів чи суб'єктів владних повноважень, а й особи, щодо якої рішення приймається. Включаючи закріплення таких гарантій як право на інформацію про факт використання відповідних систем, право на тлумачення логіки та причин остаточного рішення, з'ясування його правових наслідків і можливості оскарження, нормативна регламентація права на пояснення алгоритмічних рішень є одним із важливих кроків на шляху до встановлення належного рівня захисту прав і свобод людини в умовах цифрової трансформації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Rodà A. The COMPAS case: an educational journey for explaining fairness in AI-based applications. *CEUR Workshop Proceedings*. 2025. Vol. 3961. P. 1-9. URL: <https://ceur-ws.org/Vol-3961/paper4.pdf> (дата звернення: 14.12.2025).
2. Berk R. A. Artificial intelligence, predictive policing, and risk assessment for law enforcement. *Annual Review of Criminology*. 2021. Vol. 4 (1). P. 209-237. URL: <https://lnk.ua/zN2K7yY47> (дата звернення: 14.12.2025).
3. Берназюк Я. Штучний інтелект у правосудді: виклики алгоритмічних помилок і гарантії справедливого процесу. 2025. URL: https://court.gov.ua/storage/portal/supreme/prezentacii_2025/145_Responsible_AI_Principles_in_Justice_bernaziuk.pdf (дата звернення: 14.12.2025).
4. De Bruijn H., Warnier M., Janssen M. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government information quarterly*. 2022. Vol. 39 (2), 101666. P. 1-8. URL: <https://lnk.ua/xVm8lRyev> (дата звернення: 14.12.2025).
5. Karlan B., Kugelberg H. D. No right to an explanation. *Philosophy and Phenomenological Research*. 2025. Vol. 111. P. 137-156. URL: <https://onlinelibrary.wiley.com/doi/full/10.1111/phpr.70008> (дата звернення: 14.12.2025).
6. Kim T. W., Routledge B. R. Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. *Business Ethics Quarterly*. 2022. Vol. 32 (1). P. 75-102. URL: <https://lnk.ua/Y4Q96Ml49> (дата звернення: 14.12.2025).
7. Про захист фізичних осіб у зв'язку з опрацюванням персональних даних і про вільний рух таких даних, та про скасування Директиви 95/46/ЄС (Загальний регламент про за-

хист даних) : Регламент Європейського Парламенту і Ради (ЄС) від 27 квітня 2016 року. № 2016/679. URL: https://zakon.rada.gov.ua/laws/show/984_008-16#Text (дата звернення: 14.12.2025).

8. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised 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). URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng> (дата звернення: 14.12.2025).