Digital Ethics and Algorithmic Transparency: Challenges and Approaches to Ensuring Fairness in Automated Decision-Making within Public Administration

UDC: UDC 35.08:004.8:17(477) DOI: https://doi.org/10.15421/152520

Orlov Oleksandr
Dr.Sc., Full Prof., https://orcid.org/0000-0001-8995-7383, avorlovav@gmail.com
V. N. Karazin Kharkiv National University (Kharkiv, Ukraine)

Abstract.
Key challenges of digital ethics and algorithmic transparency in the context of implementing automated decision-making systems in public administration have been investigated. Four main categories of ethical risks have been identified: discriminatory risks associated with the reproduction and amplification of social biases through algorithmic systems; autonomy risks arising from reduced human control over important decisions; accountability risks caused by the complexity of determining responsibility in multi-level technical systems; and transparency risks concerning the opacity of algorithmic processes for citizens and civil servants.

Methods for ensuring algorithmic fairness have been systematized into three main categories. Technical methods have been analyzed, including data preprocessing to eliminate biases, fair learning algorithms using multi-objective optimization and adversarial approaches, as well as post-processing of results through calibration of decision-making thresholds. Procedural guarantees have been characterized, including mechanisms for citizen participation in system development, creation of accessible procedures for appealing algorithmic decisions, and regular auditing of algorithmic systems.

The impact of algorithmic opacity on democratic principles has been assessed, revealing three main dimensions of this problem: technical opacity due to the complexity of algorithmic models, procedural opacity due to lack of information about algorithm integration into government processes, and organizational opacity due to the complexity of responsibility distribution among different actors.

A comprehensive set of practical recommendations for implementing ethical algorithmic systems in public administration has been developed.

Keywords: digital ethics, algorithmic transparency, automated decision-making systems, public administration, algorithmic fairness, artificial intelligence, discriminatory and ethica risks, digital transformation

Цифрова етика та алгоритмічна прозорість: виклики та методи забезпечення справедливості автоматизованих рішень у державному управлінні

Орлов Олександр
Харківський національний університет імені В. Н. Каразіна (Харків, Україна)

Анотація.
У статті досліджено ключові виклики цифрової етики та алгоритмічної прозорості в контексті впровадження автоматизованих систем прийняття рішень у державному управлінні. Виявлено чотири основні категорії етичних ризиків: дискримінаційні ризики, пов'язані з відтворенням та посиленням соціальних упереджень через алгоритмічні системи; ризики автономії, що виникають через зменшення людського контролю над важливими рішеннями; ризики підзвітності, обумовлені складністю визначення відповідальності в багаторівневих технічних системах; та ризики прозорості, що стосуються непрозорості алгоритмічних процесів для громадян та державних службовців.

Систематизовано методи забезпечення алгоритмічної справедливості за трьома основними категоріями. Проаналізовано технічні методи, включаючи предобробку даних для усунення упереджень, справедливі алгоритми навчання з використанням багатоцільової оптимізації та адверсаріальних підходів, а також постобробку результатів через калібрування порогів прийняття рішень. Охарактеризовано процедурні гарантії, включаючи механізми участі громадян у розробці систем, створення доступних процедур оскарження алгоритмічних рішень та регулярний аудит алгоритмічних систем.

Оцінено вплив алгоритмічної непрозорості на демократичні принципи та виявлено три основні виміри цієї проблеми: технічну непрозорість через складність алгоритмічних моделей, процедурну непрозорість через відсутність інформації про інтеграцію алгоритмів у державні процеси та організаційну непрозорість через складність розподілу відповідальності між різними акторами.

Розроблено комплекс практичних рекомендацій для впровадження етичних алгоритмічних систем у державному управлінні.

Ключові слова: цифрова етика, алгоритмічна прозорість, автоматизовані системи прийняття рішень, публічне управління, алгоритмічна справедливість, штучний інтелект, дискримінаційні та етичні ризики, цифрова трансформація

Стаття надійшла / Article arrived: 01.05.25 Схвалено до друку / Accepted: 26.06.25

Вступ.

Сучасне державне управління переживає період безпрецедентної цифрової трансформації, характерний широким впровадженням технологій штучного інтелекту та алгоритмічних систем прийняття рішень. Від автоматизованого розподілу соціальних допомог до систем кримінального правосуддя, від алгоритмів оцінки кредитоспроможності до систем розпізнавання обличь у публічних місцях, алгоритми все частіше визначають долі громадян і формують соціальну справедливість у сучасному суспільстві (Binns, 2018; Yeung, 2018).

Ця тенденція відображає глобальну потребу в підвищенні ефективності державних послуг, зменшенні бюрократичних процедур та оптимізації використання публічних ресурсів. Алгоритмічні системи обіцяють швидкість, консистентність та об'єктивність у прийнятті рішень, що традиційно характеризувалися суб'єктивністю та непередбачуваністю людського фактору (Bullock, 2019). Однак разом із перевагами автоматизації виникають серйозні етичні дилеми, які ставлять під сумнів саму концепцію об'єктивності машинних рішень.

Непрозорість алгоритмічних процесів, потенційні упередження в даних та відсутність підзвітності створюють ризики для демократичних принципів та основних прав людини (Citron & Pasquale, 2014). Особливо гостро ці питання постають у контексті державного управління, де рішення алгоритмів можуть мати далекосяжні наслідки для життя громадян, починаючи від доступу до соціальних послуг і закінчуючи питаннями особистої свободи та безпеки.

Проблема ускладнюється тим, що сучасні системи штучного інтелекту, особливо ті, що базуються на глибокому навчанні, часто функціонують як "чорні скрині", коли навіть їх розробники не можуть повністю пояснити логіку прийняття конкретних рішень (Rudin, 2019). Це створює парадоксальну ситуацію, коли державні інституції, які традиційно мають бути підзвітними громадянам, використовують інструменти, внутрішня логіка яких залишається непрозорою.

Актуальність дослідження зумовлена зростаючою потребою в розробці етичних стандартів та методологій забезпечення прозорості алгоритмічних систем у публічному секторі. В Україні цифрова трансформація державного управління, зокрема через платформу Дія (Міністерство цифрової трансформації України, 2024), підкреслює необхідність етичних стандартів для забезпечення довіри громадян. Незважаючи на численні дискусії щодо цифрової етики на міжнародних форумах та в академічних колах, досі бракує систематичних досліджень методів практичного впровадження принципів справедливості в автоматизовані рішення, особливо в контексті специфічних вимог державного управління (Jobin et al., 2019).

Аналіз попередніх публікацій. Цифрова етика як наукова дисципліна сформувалася на перетині комп'ютерних наук, філософії та соціології наприкінці XX століття, проте особливої актуальності набула з розвитком технологій машинного навчання та штучного інтелекту. Фундаментальні принципи цифрової етики включають автономію людини, справедливість, непошкодження та пояснюваність алгоритмічних рішень. Ці принципи розроблялися як адаптація традиційних етичних концепцій до специфіки цифрових технологій.

Флоріді та колеги у своїй впливовій роботі 2018 року запропонували концепцію "етики для доброго штучного інтелекту", яка визначає етичні рамки не лише як обмеження для технологій, але й як позитивну силу для створення кращого суспільства. Вони наголошують на необхідності проактивного підходу до етики ШІ, коли етичні принципи інтегруються в процес розробки з самого початку, а не додаються постфактум (Floridi et al., 2018).

Концепція алгоритмічної справедливості розвивалася паралельно з усвідомленням потенційних ризиків автоматизованих систем. Дослідники виділяють кілька типів справедливості, кожен з яких має свої переваги та обмеження (Chouldechova, 2017). Індивідуальна справедливість передбачає однакове ставлення до схожих індивідів, проте визначення "схожості" часто є проблематичним. Групова справедливість фокусується на рівності результатів для різних демографічних груп, але може призводити до зворотної дискримінації. Процедурна справедливість зосереджується на справедливості самого процесу прийняття рішень, незалежно від результатів (Dwork et al., 2012).

Проблема алгоритмічних упереджень стала центральною в дискусіях про етику штучного інтелекту після ряду резонансних випадків дискримінації. Упередження в алгоритмічних системах можуть виникати на різних етапах їх життєвого циклу, створюючи складну мережу взаємопов'язаних проблем (Barocas & Selbst, 2016). На етапі збору даних упередження можуть з'являтися через неповноту вибірки, коли певні

групи населення недостатньо представлені в навчальних даних, або через систематичні помилки у процесі збору.

Баумер та Сілберг у своєму дослідженні 2019 року класифікують упередження як історичні, що відображають минулі дискримінаційні практики, репрезентаційні, що виникають через неповноту або спотворення даних, та агрегаційні, що пов'язані з некоректним узагальненням різнорідних груп (Baumer & Silberg, 2019). Ця класифікація допомагає зрозуміти різні джерела упереджень та розробити відповідні стратегії їх подолання.

Особливо проблематичними є випадки, коли алгоритми не лише відтворюють, але й підсилюють існуючі соціальні нерівності. Класичним прикладом є система COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), що використовувалася в американському правосудді для оцінки ризику рецидиву. Журналістське розслідування ProPublica показало, що система демонструвала расові упередження, помилково класифікуючи афроамериканських підсудних як високоризикових удвічі частіше, ніж білих (Angwin et al., 2016).

Проблема ускладнюється тим, що упередження можуть бути неявними та проявлятися лише через статистичний аналіз результатів роботи системи. Більше того, спроби виправити одні типи упереджень можуть призводити до появи інших, створюючи складні етичні дилеми щодо пріоритетності різних видів справедливості (Kleinberg et al., 2017).

Прозорість алгоритмів є багатовимірним концептом, що включає доступність інформації про алгоритм, зрозумілість його роботи та можливість перевірки рішень. Діакопулос у своїй роботі 2016 року розрізняє процедурну прозорість, що передбачає розкриття логіки алгоритму та його параметрів, та результативну прозорість, що фокусується на поясненні конкретних рішень для окремих випадків (Diakopoulos, 2016).

Концепція пояснювального штучного інтелекту виникла як відповідь на проблему "чорної скрині" складних алгоритмічних моделей, особливо нейронних мереж глибокого навчання. Гудман та Флаксман у своєму дослідженні 2017 року наголошують на правовому аспекті пояснюваності, зокрема у контексті європейського регулювання, що надає громадянам право на пояснення автоматизованих рішень (Goodman & Flaxman, 2017).

Методи пояснення включають локальні пояснення, що фокусуються на конкретних рішеннях та показують, які фактори вплинули на результат у конкретному випадку, та глобальні пояснення, що описують загальну логіку роботи моделі (Ribeiro et al., 2016). Кожен підхід має свої переваги та обмеження, і вибір відповідного методу залежить від контексту використання та потреб користувачів.

Важливо відзначити, що пояснюваність не є універсальним розв'язанням всіх етичних проблем. Надмірна деталізація пояснень може призводити до інформаційного перевантаження користувачів, тоді як спрощені пояснення можуть бути не лише неточними, але й оманливими (Wachter et al., 2017). Крім того, сам факт надання пояснення не гарантує справедливості або етичності рішення.

Регулювання алгоритмічних систем стало важливою темою для урядів та міжнародних організацій. Європейський Союз займає провідну позицію в цій сфері, розробивши Загальний регламент про захист даних (GDPR) та працюючи над Актом про штучний інтелект (Veale & Edwards, 2018). Ці документи встановлюють правові рамки для використання алгоритмічних систем та надають громадянам певні права щодо автоматизованого прийняття рішень.

В Сполучених Штатах підхід до регулювання є більш фрагментованим, з різними ініціативами на федеральному та місцевому рівнях. Деякі міста, такі як Нью-Йорк та Сан-Франциско, прийняли місцеві закони про алгоритмічну підзвітність, що вимагають від державних установ розкривати інформацію про використовувані алгоритмічні системи (Kroll et al., 2017).

Міттельштадт у своєму дослідженні 2016 року аналізує різні підходи до аудиту алгоритмічних систем та наголошує на необхідності розробки стандартизованих процедур перевірки (Mittelstadt, 2016). Він виділяє технічний аудит, що фокусується на точності та справедливості алгоритмів, та соціальний аудит, що розглядає ширший вплив алгоритмічних систем на суспільство.

Огляд публікацій демонструє зростаючу увагу світової наукової спільноти до етичних викликів ШІ, що підтверджує актуальність обраної мети дослідження. Аналіз наукових джерел дозволяє визначити, які аспекти цифрової етики та алгоритмічної прозорості вже досліджені, а які залишаються недостатньо вивченими, особливо в контексті державного управління. Аналіз літератури показує брак конкретних рекомендацій для пост-радянських

країн, що обґрунтовує необхідність розробки практичних рекомендацій у сфері публічного управління України. Метою статті є комплексний аналіз викликів цифрової етики та алгоритмічної прозорості у сфері державного управління з розробкою практичних рекомендацій щодо етичного впровадження автоматизованих систем прийняття рішень.

Для досягнення поставленої мети визначено такі завдання:

1. Ідентифікувати та систематизувати основні категорії етичних ризиків, пов'язаних з використанням алгоритмічних систем у державному управлінні.
2. Проаналізувати методи забезпечення алгоритмічної справедливості та класифікувати їх за технічними, процедурними та правовими критеріями.
3. Оцінити вплив алгоритмічної непрозорості на реалізацію демократичних принципів у публічному управлінні.
4. Дослідити міжнародний досвід регулювання етичних аспектів використання штучного інтелекту в державному секторі.
5. Розробити комплекс стратегічних, оперативних та правових рекомендацій для впровадження етичних алгоритмічних систем у державному управлінні України.

Методологія дослідження. Дослідження базується на змішаній методології, що поєднує кілька взаємодоповнюючих підходів для забезпечення комплексного розуміння проблеми цифрової етики та алгоритмічної прозорості в державному управлінні. Методологічний плюралізм обумовлений складністю досліджуваного явища, що знаходиться на перетині технологічних, соціальних, політичних та правових аспектів.

Аналітична частина дослідження включає систематичний огляд наукових публікацій за період 2018-2024 років у провідних міжнародних базах даних, включаючи Scopus, Web of Science, IEEE Xplore та ACM Digital Library. Пошук здійснювався за комбінацією ключових термінів англійською мовою, таких як "algorithmic transparency", "AI ethics", "government algorithms", "algorithmic fairness", "explainable AI" та "automated decision-making". Загалом було проаналізовано понад 150 наукових публікацій, з яких 89 були включені до детального аналізу після застосування критеріїв релевантності та якості.

Критеріями включення публікацій до аналізу служили їх фокус на етичних аспектах використання алгоритмічних систем у публічному секторі, наявність емпіричних даних або конкретних методологічних пропозицій, а також публікація в рецензованих наукових виданнях. Особлива увага приділялася роботам, що розглядають практичні аспекти впровадження етичних принципів у реальних алгоритмічних системах.

Емпірична частина дослідження базується на аналізі практичних кейсів впровадження алгоритмічних систем у державних установах різних країн. Було детально вивчено досвід Нідерландів із системою автоматизованого виявлення шахрайства в соціальному забезпеченні, яка призвела до скандалу з несправедливими звинуваченнями тисяч сімей (Peeters & Widlak, 2018). Також проаналізовано естонську платформу електронного уряду, що використовує алгоритмічні системи для надання адміністративних послуг (Anthes, 2015), та фінську систему розподілу освітніх ресурсів на основі алгоритмічних рішень.

Додатково було проведено аналіз документів та звітів міжнародних організацій, включаючи ОЕСР, Ради Європи, ЮНЕСКО та Світового банку, щодо рекомендацій з етичного використання штучного інтелекту в державному управлінні. Ці документи надають важливий контекст для розуміння міжнародних тенденцій та стандартів у сфері алгоритмічної етики (OECD, 2019; Council of Europe, 2018).

Методологія також включає компаративний аналіз регулятивних підходів до алгоритмічної прозорості в різних юрисдикціях. Особлива увага приділялася порівнянню європейського підходу, що базується на правах людини та сильному регулюванні, з американським підходом, що більше покладається на саморегулювання та ринкові механізми.

Результати.

Аналіз проблем цифрової етики в державному управлінні. Впровадження алгоритмічних систем у державне управління породжує широкий спектр етичних ризиків, які можна систематизувати за кількома ключовими категоріями. Дискримінаційні ризики становлять, можливо, найбільш серйозну загрозу для принципів демократичного управління та рівності перед законом. Ці ризики пов'язані з потенційним несправедливим ставленням до окремих груп громадян на основі їх демографічних характеристик, соціально-економічного статусу або інших факторів (Selbst et al., 2019).

Алгоритми можуть відтворювати історичні упередження, закодовані в навчальних

даних, що містять інформацію про минулі дискримінаційні практики. Наприклад, якщо в минулому певні групи мали менший доступ до освіти або трудових можливостей, алгоритм може інтерпретувати це як свідчення їх нижчої "придатності" для отримання освітніх грантів або працевлаштування в державному секторі. Більше того, алгоритми можуть створювати нові форми дискримінації через некоректну інтерпретацію даних або використання pro-ху-змінних, що корелюють з захищеними характеристиками (Barocas & Selbst, 2016).

Ризики автономії виникають через зменшення людського контролю над важливими рішеннями, що впливають на життя громадян. Автоматизація державних процесів може призвести до втрати можливості індивідуального розгляду справ та врахування особливих обставин, що не передбачені алгоритмічною моделлю. Це особливо проблематично в ситуаціях, коли стандартні процедури не підходять для конкретного випадку або коли є потреба в людському співчутті та розумінні (Raso et al., 2018).

Проблема ускладнюється тим, що державні службовці можуть поступово втрачати навички критичного аналізу та прийняття рішень, покладаючись на алгоритмічні рекомендації. Це може призвести до ситуації, коли люди стають просто виконавцями алгоритмічних рішень, не розуміючи їх логіки та не маючи можливості або бажання їх оспорювати (Green & Chen, 2019).

Ризики підзвітності пов'язані з труднощами визначення відповідальності за алгоритмічні рішення в складних організаційних структурах державного управління. Складність сучасних систем штучного інтелекту ускладнює встановлення причинно-наслідкових зв'язків та розподіл відповідальності між різними акторами, включаючи розробників алгоритмів, державних службовців, що їх використовують, та керівників, що приймають рішення про їх впровадження (Ananny & Crawford, 2018).

Традиційні механізми підзвітності в державному управлінні можуть виявитися неефективними в контексті алгоритмічних систем. Парламентський контроль, судовий перегляд та громадський нагляд можуть бути ускладнені технічною складністю алгоритмічних систем та відсутністю відповідної експертизи у контролюючих органів (Coglianese & Lehr, 2017).

Ризики прозорості включають непрозорість алгоритмічних процесів не лише для громадян, але й для державних службовців, що їх використовують. Це підриває демократичні принципи відкритості та підзвітності влади, створюючи ситуацію, коли рішення приймаються за допомогою інструментів, логіка яких є недоступною для розуміння (Pasquale, 2015).

Державний сектор має особливості, що ускладнюють впровадження етичних алгоритмічних систем порівняно з приватним сектором. Обов'язковість взаємодії з державними алгоритмічними системами створює особливу етичну відповідальність для урядових інституцій . На відміну від комерційних сервісів, які громадяни можуть обирати або відмовлятися від їх використання, державні алгоритми часто є єдиним способом отримання необхідних послуг або дотримання законних вимог (Citron, 2007).

Ця обов'язковість означає, що громадяни не можуть "проголосувати ногами" проти несправедливих або упереджених алгоритмічних систем, як вони могли б зробити з комерційними продуктами. Це створює особливу відповідальність для державних інституцій щодо забезпечення справедливості та етичності своїх алгоритмічних систем.

Масштаб впливу державних алгоритмів значно перевищує вплив більшості комерційних систем. Державні алгоритми можуть впливати на мільйони громадян одночасно, що означає, що навіть невеликі упередження або помилки можуть мати катастрофічні наслідки на рівні суспільства. Скандал з голландською системою виявлення шахрайства в дитячих допомогах, що несправедливо позначила тисячі сімей як шахраїв, демонструє масштаб потенційної шкоди (Peeters & Widlak, 2018). В Україні цифрова трансформація місцевого врядування стикається з викликами, пов’язаними з браком технічної експертизи та нерівномірним доступом до цифрових сервісів, що може посилювати алгоритмічні упередження. Наприклад, у віддалених регіонах обмежений доступ до інтернету та низька цифрова грамотність можуть призводити до недостатньої репрезентації даних про ці громади, що впливає на якість алгоритмічних рішень.

Складність правового регулювання створює додаткові виклики для державних алгоритмічних систем. Ці системи повинні відповідати численним правовим нормам, включаючи конституційні принципи рівності та справедливості, адміністративне право, законодавство про захист персональних даних та галузеві нормативні акти. Узгодження всіх цих вимог з технічними можливостями алгоритмічних систем є складним завданням (Engstrom & Ho, 2020). Платформа Дія, що

забезпечує доступ до державних послуг, демонструє успіхи цифровізації, але потребує прозорих алгоритмів для підтримки довіри (Міністерство цифрової трансформації України, 2024).

Розробка етичних рамок для використання штучного інтелекту в українському державному управлінні має враховувати локальні виклики, такі як потреба в гармонізації з європейськими стандартами та уникнення дискримінації в автоматизованих системах (Карпенко, 2019). В Україні етичні аспекти використання ШІ потребують чітких нормативних рамок, щоб забезпечити прозорість і підзвітність, зокрема в державному управлінні та наукових дослідженнях, де ШІ може впливати на об'єктивність і достовірність даних (Бердо та ін., 2023; Єфіменко, 2024)

Політична сенситивність державних алгоритмічних рішень додає ще один рівень складності. Алгоритмічні рішення в державному секторі часто мають політичні імплікації та можуть впливати на довіру громадян до влади. Помилки або упередження в державних алгоритмах можуть стати предметом політичних дебатів та впливати на електоральні перспективи правлячих партій.

Це створює складну дилему між потребою в інновації та необхідністю уникнення політичних ризиків. З одного боку, уряди мають стимули впроваджувати нові технології для підвищення ефективності та якості державних послуг. З іншого боку, вони мають бути надзвичайно обережними щодо потенційних негативних наслідків, які можуть підірвати довіру громадян (Карпенко 2019).

Алгоритмічна непрозорість у державному управлінні створює фундаментальний виклик для демократичних принципів. У демократичному суспільстві громадяни мають право знати, як приймаються рішення, що впливають на їх життя. Проте сучасні алгоритмічні системи, особливо ті, що базуються на машинному навчанні, часто функціонують як "чорні скрині", де навіть їх розробники не можуть повністю пояснити логіку конкретних рішень (Burrell, 2016).

Ця непрозорість має кілька вимірів. Технічна непрозорість виникає через складність алгоритмічних моделей, особливо нейронних мереж з великою кількістю параметрів. Навіть якщо код алгоритму є відкритим, розуміння його роботи може вимагати високого рівня технічної експертизи, недоступної більшості громадян та навіть державних службовців. Більш того, вона має принциповий характер. Однією з ключових причин алгоритмічної непрозорості великих мовних моделей (LLM) є змагальний принцип тренування, втілений у взаємодії Генератора та Дискримінатора в генеративно-змагальних мережах (GAN). Генератор, отримуючи випадковий шум, створює синтетичні дані, намагаючись імітувати реальні, тоді як Дискримінатор оцінює їх правдоподібність, розрізняючи справжні та фальшиві зразки (Yu et al., 2017). Цей процес, відомий як мінімаксна гра, оптимізує Генератор для створення даних, які Дискримінатор не може ідентифікувати як фальшиві, і Дискримінатор — для точного розпізнавання (Edwards & Storkey, 2016). Математично це виражається як:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}}[\log D(x)] + E_{z\sim p_z}\left[\log\left(1-D(G(z))\right)\right],$$

де G — Генератор, D — Дискримінатор, x – реальні дані, z – шум, D(x) – ймовірність, що x справжнє. У процесі тренування досягається рівновага Неша, коли Генератор створює дані, які Дискримінатор не може відрізнити від реальних (ймовірність ~0.5). У рівновазі Неша, коли ймовірність розпізнавання становить ~0.5, Генератор продукує дані, що неможливо відрізнити від реальних (Burrell, 2016). Ця невідрізнюваність наближає ШІ до проходження тесту Тьюрінга, де машина імітує людську поведінку настільки, що стає нерозпізнаваною, але водночас ускладнює пояснення її рішень, поглиблюючи непрозорість (Turing, 1950). У державному управлінні, де прозорість є основою довіри, така "чорна скринька" створює етичні виклики.

Для подолання непрозорості пропонуються комплексні методи. Технічно, методи пояснювального штучного інтелекту (XAI), як-от LIME та SHAP, дозволяють ідентифікувати фактори, що впливають на рішення моделей, частково розкриваючи змагальну логіку (Ribeiro et al., 2016). Організаційно, стандарти документування тренувальних даних і архітектури моделей підвищують підзвітність, що є актуальним для української платформи Дія, де довіра громадян є пріоритетом. Процедурна прозорість досягається через публічні консультації, які залучають громадян до оцінки алгоритмів (Young et al., 2019). Хоча повна прозорість залишається недосяжною через складність змагальних процесів, ці методи створюють основу для етичного використання ШІ в державному секторі, сприяючи справедливості та довірі.

Процедурна непрозорість пов'язана з

відсутністю чіткої інформації про те, як алгоритмічні системи інтегровані в державні процеси, хто приймає рішення про їх використання та як вони взаємодіють з людьми--операторами. Громадяни часто не знають, чи їх справа розглядається людиною, алгоритмом, чи їх комбінацією (Kemper & Kolkman, 2019).

Організаційна непрозорість виникає через складну структуру сучасних державних інституцій, де відповідальність за алгоритмічні рішення може бути розподілена між різними відомствами, підрядниками та техніками. Це ускладнює ідентифікацію конкретних осіб або органів, відповідальних за алгоритмічні рішення.

Методи забезпечення алгоритмічної справедливості. Технічні підходи до забезпечення алгоритмічної справедливості розвивалися як відповідь на виявлення упереджень у машинному навчанні та фокусуються на модифікації алгоритмічних процесів для досягнення більш справедливих результатів. Предобробка даних являє собою комплекс методів, спрямованих на виявлення та корекцію упереджень у навчальних даних ще до тренування алгоритмічних моделей (Kamiran & Calders, 2012).

Одним з найпоширеніших підходів є репалансування навчальних даних для забезпечення пропорційного представлення різних груп. Проте простий перерозподіл може призвести до втрати важливої інформації або створення штучних патернів, що не відображають реальності. Більш складні техніки включають синтетичну генерацію даних за допомогою генеративних моделей, що дозволяє створювати реалістичні приклади для недопредставлених груп (Chawla et al., 2002).

Методи видалення чутливих атрибутів спрямовані на усунення з навчальних даних інформації про характеристики, що можуть призводити до дискримінації, такі як стать, раса або релігія. Проте ця техніка має обмеження, оскільки алгоритми можуть використовувати proxy-змінні, що корелюють з видаленими атрибутами. Наприклад, поштовий індекс може бути індикатором расового складу району, навіть якщо раса не вказана явно в даних (Pedreshi et al., 2008).

Справедливі алгоритми навчання представляють більш складний підхід, що передбачає модифікацію самого процесу тренування моделей для одночасної оптимізації точності та справедливості. Ці методи включають додавання обмежень на справедливість до функції втрат, використання регуляризаційних термів та багатоцільову оптимізацію, що дозволяє балансувати між різними метриками ефективності (Zafar et al., 2017).

Адверсаріальні підходи до справедливого навчання використовують концепцію змагальних мереж, де одна модель намагається зробити точні передбачення, а інша намагається виявити упередження в цих передбаченнях. Цей підхід може бути особливо ефективним для виявлення прихованих упереджень, що не очевидні при традиційному аналізі (Edwards & Storkey, 2016).

Постобробка результатів дозволяє корегувати вихідні дані алгоритмічних моделей для досягнення бажаних показників справедливості без необхідності перетренування моделей. Методи калібрування можуть налаштовувати пороги прийняття рішень для різних груп, щоб забезпечити рівні показники помилок. Техніки перерозподілу рішень можуть змінювати результати для досягнення статистичної паритетності між групами.

Проте всі технічні методи мають свої обмеження та компроміси. Підвищення справедливості часто призводить до зниження загальної точності моделі, створюючи дилему між ефективністю та етичністю. Більше того, різні метрики справедливості можуть суперечити одна одній, роблячи неможливим одночасне досягнення всіх видів справедливості для різних груп, забезпечуючи однакові показники помилкових позитивних або негативних результатів (Hardt et al., 2016).

Організаційні підходи до забезпечення алгоритмічної справедливості фокусуються на створенні інституційних механізмів та процедур, що забезпечують етичне використання алгоритмічних систем у державному управлінні. Створення етичних комітетів та наглядових рад з алгоритмічної етики стає дедалі поширенішою практикою в урядових установах (Jobin et al., 2019).

Ці комітети мають включати представників різних дисциплін - від технічних спеціалістів до правників, соціологів та представників громадянського суспільства. Мультидисциплінарний підхід забезпечує комплексну оцінку етичних ризиків та розробку збалансованих рішень, що враховують технічні можливості, правові вимоги та соціальні наслідки.

Розробка внутрішніх стандартів та політик використання алгоритмічних систем є критично важливою для забезпечення консистентності етичних підходів у різних підрозділах державної організації. Ці стандарти повинні визначати процедури оцінки етичних ризиків, вимоги

до документування алгоритмічних рішень та протоколи реагування на виявлені проблеми (Reisman et al., 2018).

Навчання персоналу з питань алгоритмічної етики є невід'ємною частиною організаційних заходів. Державні службовці, що працюють з алгоритмічними системами, повинні розуміти основні етичні принципи, вміти виявляти потенційні упередження та знати процедури ескалації проблемних ситуацій.

Процедурні гарантії включають механізми участі громадян у процесах розробки та впровадження алгоритмічних систем. Публічні консультації на етапі планування дозволяють виявити потенційні проблеми та врахувати інтереси різних груп населення ще до початку розробки системи (Young et al., 2019).

Створення механізмів подання скарг та оскарження алгоритмічних рішень є критично важливим для забезпечення справедливості. Ці механізми повинні бути доступними, зрозумілими для громадян та забезпечувати ефективний перегляд спірних рішень кваліфікованими спеціалістами.

Регулярний аудит алгоритмічних систем дозволяє виявляти та усувати упередження, що можуть розвиватися з часом через зміни в даних або контексті використання. Аудит повинен включати як технічну перевірку точності та справедливості алгоритмів, так і оцінку їх соціального впливу (Raji et al., 2020).

Рекомендації для практичного впровадження. Розробка національної стратегії етичного використання штучного інтелекту в державному управлінні повинна стати пріоритетом для урядів, що прагнуть скористатися перевагами цифрової трансформації, зберігаючи при цьому демократичні цінності та права громадян. В Україні розробка такої стратегії має враховувати досягнення цифровізації, але й долати виклики, як-от брак координації між відомствами.

Створення спеціалізованого органу з нагляду за алгоритмічними системами може забезпечити централізований контроль та координацію зусиль різних відомств. Такий орган повинен мати відповідні повноваження, ресурси та експертизу для ефективного виконання своїх функцій.

Інвестиції в освіту та підготовку кадрів є критично важливими для успішного впровадження етичних алгоритмічних систем. Це включає як підготовку технічних фахівців з етики ШІ, так і навчання державних службовців основам алгоритмової грамотності.

На оперативному рівні рекомендується впровадження обов'язкової оцінки впливу на алгоритмічну справедливість (Algorithmic Impact Assessment) для всіх нових алгоритмічних систем у державному секторі. Ця оцінка повинна включати аналіз потенційних упереджень, оцінку ризиків для різних груп населення та план заходів з мітигації виявлених проблем.

Створення реєстрів алгоритмічних систем, що використовуються в державному управлінні, підвищить прозорість та дозволить громадянам знати, які алгоритми впливають на їх життя. Ці реєстри повинні включати базову інформацію про призначення систем, типи рішень, що приймаються, та контактну інформацію для звернень.

Розробка стандартизованих форматів пояснення алгоритмічних рішень допоможе забезпечити зрозумілість для громадян. Пояснення повинні бути адаптовані до рівня технічної грамотності цільової аудиторії та включати інформацію про ключові фактори, що вплинули на рішення.

Модернізація адміністративного законодавства для врахування специфіки алгоритмічних рішень є необхідною умовою для створення адекватної правової бази. Розробка спеціального законодавства про алгоритмічну прозорість у державному секторі може забезпечити більш детальне регулювання цієї сфери. Таке законодавство повинне встановлювати мінімальні стандарти прозорості, вимоги до документування та процедури. В Україні законодавчі ініціативи щодо ШІ можуть спиратися на етичні принципи, адаптовані до локального контексту.

Створення правових механізмів відповідальності за шкоду, завдану алгоритмічними рішеннями, забезпечить адекватний захист прав громадян. Це включає як індивідуальну відповідальність посадових осіб, так і інституційну відповідальність державних органів.

Висновки.

Дослідження цифрової етики та алгоритмічної прозорості в державному управлінні виявляє складну мережу взаємопов'язаних викликів, що потребують комплексного міждисциплінарного підходу для їх вирішення. Впровадження алгоритмічних систем у публічний сектор несе значні переваги в плані ефективності та консистентності, проте водночас створює нові ризики для демократичних принципів та основних прав людини.

Основні виклики включають потенційну дискримінацію через алгоритмічні упередження, зменшення людської автономії

через надмірну автоматизацію, проблеми підзвітності в складних технічних системах та непрозорість алгоритмічних процесів для громадян та навіть державних службовців. Ці виклики ускладнюються специфічними характеристиками державного сектору, включаючи обов'язковість взаємодії, масштаб впливу та політичну сенситивність рішень.

Методи забезпечення алгоритмічної справедливості охоплюють технічні підходи до корекції упереджень у даних та алгоритмах, організаційні механізми контролю та нагляду, а також процедурні гарантії участі громадян та оскарження рішень. Ефективність цих методів залежить від їх комплексного впровадження та адаптації до специфічних умов кожної юрисдикції.

Практичні рекомендації включають розробку національних стратегій етичного використання ШІ, створення спеціалізованих наглядових органів, впровадження обов'язкової оцінки впливу на алгоритмічну справедливість та модернізацію правової бази. Особлива увага приділяється необхідності інвестицій в освіту та підготовку кадрів, здатних працювати на перетині технологій та етики.

Перспективи подальших досліджень включають розробку більш складних метрик справедливості, що враховують контекстуальні особливості різних сфер державного управління, дослідження довгострокових соціальних наслідків автоматизації державних процесів та аналіз ефективності різних регулятивних підходів у різних правових системах.

Успішне впровадження етичних алгоритмічних систем у державне управління потребує тривалих зусиль та співпраці між технічними спеціалістами, політиками, правниками та громадянським суспільством. Лише через такий мультистейкхолдерський підхід можна забезпечити, що цифрова трансформація державного управління сприятиме зміцненню демократичних цінностей та підвищенню добробуту громадян.

REFERENCES

Ananny, M., & Crawford, K. (2018). Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. *New Media & Society,* 20(3), 973-989. https://doi.org/10.1177/1461444816676645.

Angwin, Julia, Jeff Larson, Surya Mattu, & Lauren Kirchner. (2016). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Anthes, G. (2015). Estonia: A Model for E-Government. *Communications of the ACM* 58, 6, 18-20. https://doi.org/10.1145/2749414.

Barocas, Solon, & Andrew D. Selbst. (2016). «Big Data's Disparate Impact.» *California Law Review* 104(3), 671-732. https://doi.org/10.15779/Z38BG31.

Baumer, E., & Silberg, J. (2019). «Re -shaping AI to Serve Humans, Not Shareholders. In *Towards a Human-AI Ecosystem*, 1-15. Partnership on AI.

Berdo, R. S., Rasiun, V. L., & Velychko, V. A. (2023). Shtuchnyi intelekt ta ioho vplyv na etychni aspekty naukovykh doslidzhen v ukrainskykh zakladakh osvity. *Akademichni Vizii*, (22). https://doi.org/10.5281/zenodo.8174388

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research* 81, 149-59. http://proceedings.mlr.press/v81/binns18a.html.

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. *The American Review of Public Administration* 49(7), 751-61. https://doi.org/10.1177/0275074019856123.

Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3(1), 1-12. https://doi.org/10.1177/2053951715622512.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, & W. Philip Kegelmeyer. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16, 321-57. https://doi.org/10.1613/jair.953.

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5(2), 153-63. https://doi.org/10.1089/big.2016.0047.

Citron, D. K. (2007). Technological Due Process. *Washington University Law Review,* 85(6), 1249-313. https://openscholarship.wustl.edu/law_lawreview/vol85/iss6/2.

Citron, D. K., & Pasquale, F. (2014). The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 89(1), 1-33. https://digitalcommons.law.uw.edu/wlr/vol89/iss1/2.

Coglianese, C., & Lehr, D. (2017). Regulating by Robot: Administrative Decision Making in the Machine-Learning Era. *Georgetown Law Journal* 105(5), 1147-223. https://georgetownlawjournal.org/articles/250/regulating-by-robot/pdf.

Council of Europe. (2018). *Declaration by the Committee of Ministers on the Manipulative Capabilities of Algorithmic Processes*. https://rm.coe.int/declaration-en/16808fa044

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62. https://doi.org/10.1145/2844110

Durman, P., & Tokhtarova, I. (2023). Vyklyky tsyfrovoi transformatsii v ukrainskomu derzhavnomu upravlinni. *Skhidnoievropeiskyi zhurnal derzhavnoi polityky*, *9*(2), 112–130.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). ACM. https://doi.org/10.1145/2090236.2090255

Edwards, H., & Storkey, A. (2016). *Censoring representations with an adversary*. (arXiv preprint arXiv:1511.05897). https://arxiv.org/abs/1511.05897

Engstrom, D. F., & Ho, D. E. (2020). Algorithmic accountability in the administrative state. *Yale Journal on Regulation*, *37*(3), 800–854. https://digitalcommons.law.yale.edu/yjreg/vol37/iss3/2

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a 'right to explanation. *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 90–99). ACM. https://doi.org/10.1145/3287560.3287563

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315-3323). https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Karpenko, Yu. V. (2019). Etychni pryntsypy zastosuvannia shtuchnoho intelektu v publichnomu upravlinni. *Visnyk Natsionalnoi akademii derzhavnoho upravlinnia pry Prezydentovi Ukrainy. Seriia: Derzhavne upravlinnia*, *(4)*, 93–97. http://nbuv.gov.ua/UJRN/vnaddy_2019_4_15

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, *22*(14), 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference* (pp. 43:1–43:23). https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(3), 633-705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

Ministerstvo tsyfrovoi transformatsii Ukrainy. (2024). *Rezultaty tsyfrovoi transformatsii v rehionakh Ukrainy za 2023 rik*. https://www.kmu.gov.ua/news/rezultaty-tsyfrovoi-transformatsii-v-rehionakh-ukrainy-za-2023-rik

Mittelstadt, B. (2016). Auditing for transparency in content personalization systems. *International Journal of Communication*, *10*, 4991–5002. https://ijoc.org/index.php/ijoc/article/view/6267

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. (OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 560–568). ACM. https://doi.org/10.1145/1401890.1401959

Peeters, R., & Widlak, A. (2018). The digital cage: Administrative exclusion through information architecture. *Information, Communication & Society*, *21*(11), 1644–1659. https://doi.org/10.1080/1369118X.2018.1518468

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*(pp. 33–44). ACM. https://doi.org/10.1145/3351095.3372873

Raso, J., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). *Artificial intelligence & human judgment: Risk assessment in the criminal justice system*. Berkman Klein Center for Internet & Society. https://dash.harvard.edu/handle/1/37342603

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). *Algorithmic impact assessments: A practical framework for public agency accountability*. AI Now Institute. https://ainowinstitute.org/aiareport2018.pdf

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). ACM. https://doi.org/10.1145/3287560.3287598

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, *34*(2), 398–404. https://doi.org/10.1016/j.clsr.2017.10.002

Vyshnevskyi, O., & Kniaziev, S. (2024). Tsyfrova transformatsiia v Ukraini: dosiahnennia ta vyklyky. *Ukrainskyi zhurnal derzhavnoho upravlinnia*, *10*(1), 88–104.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Yefimenko, I. V. (2024). Etyka shtuchnoho intelektu. *Perspektyvy ta Innovatsii Nauky (Seriia Psykholohiia)*, *(7)*, 731–738. https://doi.org/10.52058/2786-4952-2024-7(41)-731-738

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, *12*(4), 505-523. https://doi.org/10.1111/rego.12158

Young, M., Magassa, L., & Friedman, B. (2019). Toward inclusive tech policy design: A method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, *21*(2), 89-103. https://doi.org/10.1007/s10676-019-09497-z

Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (pp. 2852-2858). https://doi.org/10.1609/aaai.v31i1.10804

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics* (pp. 962-970). http://proceedings.mlr.press/v54/zafar17a.html