УДК 351:004.8
DOI https://doi.org/10.32782/TNU-2663-6468/2025.5/18

Шишкін С. О.
Університет митної справи та фінансів

ДОСВІД УПРОВАДЖЕННЯ ШІ В СИСТЕМУ СОЦІАЛЬНОГО ЗАБЕЗПЕЧЕННЯ В США ЯК ІНСТРУМЕНТУ МІНІМІЗАЦІЇ РИЗИКІВ

У статті здійснено аналіз природи та наслідків помилок, що виникають при застосуванні систем штучного інтелекту (ШІ) в публічному управлінні. На прикладі детального кейсу модернізації Управління соціального забезпечення США (SSA) проаналізована успішна модель «ШІ-асистента», що побудована на принципі «людини в контурі управління» (human-in-the-loop) як ключового інструменту мінімізації ризиків.

У статті запропоновано комплексну класифікацію потенційних помилок, а саме: помилки першого роду («відмовити нужденному»), що несуть пряму загрозу правам людини; помилки другого роду («заплатити шахраю»), що підривають фінансову стабільність та довіру до державних інституцій; «Приховані» помилки системної упередженості (algorithmic bias), що виникають з історичних даних та здатні відтворювати соціальну нерівність під маскою об'єктивності. Особливу увагу приділено найбільш вразливій ланці системи – взаємодії на стику «людина-алгоритм». Проаналізовано такі когнітивні пастки, як «автоматична упередженість» (automation bias) та проблема «чорної скриньки» (black box), які здатні нівелювати будь-які формальні запобіжники та перетворити людський контроль на ритуальну дію. Доведено, що ефективне та безпечне впровадження ШІ є не стільки технологічним, скільки комплексним інституційним викликом, що вимагає побудови багаторівневої архітектури захисту, яка поєднує процедурні гарантії, технічну прозорість (XAI) та глибоку роботу з психологією прийняття рішень. Такий підхід дозволить перетворити ШІ на надійного партнера держави, запобігаючи його трансформації у неконтрольованого «цифрового левіафана», що загрожує демократичним інститутам.

Ключові слова: *публічне управління, прийняття рішень, штучний інтелект, феномен автоматичної упередженості, пояснюваний ШІ, людський контроль, регулювання ШІ, адміністративні послуги.*

Постановка проблеми. Сучасне публічне управління стикається із двома фундаментальними викликами: соціально-економічні системи стають дедалі складнішими, а громадяни вимагають більш якісних, швидких та прозорих державних послуг. Бюрократичні моделі, побудовані на ієрархії та паперовому документообігу, вже не здатні ефективно вирішувати ці завдання. Тому цифрова трансформація з альтернативного підходу перетворилася на обов'язкову умову життєздатності та ефективності державного управління. Для України проблему актуалізує розпорядження КМУ «Про схвалення Концепції розвитку штучного інтелекту в Україні» від 2 грудня 2020 р. № 1556-р та «Біла книга «Регулювання штучного інтелекту: український контекст та рекомендації»», що була презентована Міністерством цифрової трансформації України на початку 2024 року.

Останнє десятиліття показало, що ключовим технологічним двигуном цифровізації став штучний інтелект (далі - ШІ). Якщо раніше цифрова трансформація переважно полягала у переході державних послуг в онлайн та створенні електронних реєстрів, то сьогодні вона охоплює глибокі зміни у способах прийняття управлінських рішень. Завдяки аналізу великих даних (big data), управління переходить від реактивних відповідей на вже наявні проблеми до проактивного передбачення та попередження ризиків. Наприклад, ШІ успішно використовується для оптимізації міського трафіку, прогнозування епідемічних загроз, боротьби з податковим шахрайством та персоналізації освітніх послуг [6].

Водночас із розширенням сфер застосування ШІ зростає й усвідомлення пов'язаних із цим ризиків. Впровадження алгоритмічних систем у таких чутливих сферах, як соціальне забезпечення, охорона здоров'я чи правосуддя, викликає серйозні технічні, етичні та правові питання. Серед них — проблеми надійності та безпеки

алгоритмів, упередженість систем («algorithmic bias») та невирішеність питань відповідальності за помилки «машинних рішень» [7].

Основна проблема, яка постає перед урядами, полягає в розриві між швидким технологічним розвитком і значно повільнішою інституційною адаптацією до нових умов. Особливо гостро це питання стоїть в Україні, де до загальносвітових викликів додається специфіка країни, що перебуває у стані війни, активно проводить структурні реформи та рухається шляхом європейської інтеграції. В цих умовах потреба у швидкому, але водночас обережному та виваженому впровадженні ШІ стає нагальною.

Аналіз останніх досліджень і публікацій. Питання впровадження ШІ у сферу публічного управління викликає значний інтерес у зарубіжних і вітчизняних науковців. Ця проблема висвітлена у працях українських дослідників О.О. Подворнюк, Н.В. Поліщук [1], О. Воронова, О. Остапенка, В. Яценка [2], О. Оболенського, В. Косицької, А. Рвача [3], Т.С. Ярового [4] та інших. Вони розглядають як загальні можливості та ризики застосування ШІ, так і специфічні виклики для України, зокрема проблеми якості даних, необхідність підготовки кадрів та адаптації законодавства. Незважаючи на наявність певних досліджень з визначеної проблеми, вона потребує більш глибокого осмислення і трактування.

Постановка завдання. Метою статті є визначення природи та наслідків алгоритмічних помилок при впровадженні штучного інтелекту в публічне управління. Для досягнення цієї мети поставлено завдання щодо конкретизації проблем, що виникають на стику взаємодії «людина-машина», та розробки типології ризиків, що включає як технічні, так і когнітивно-психологічні аспекти. На основі аналізу міжнародного досвіду, зокрема кейсу SSA, робота спрямована на формування комплексного розуміння інституційних та технологічних запобіжників, необхідних для побудови безпечних та ефективних систем ШІ.

Дослідження ґрунтується на поєднанні загальнонаукових методів (аналіз, синтез, узагальнення) та спеціальних методів. Ключовим став порівняльний аналіз, реалізований через кейс-стаді (case study) системи соціального забезпечення США. Для систематизації алгоритмічних помилок та ризиків було застосовано метод класифікації. Теоретичною основою роботи стали системний та інституційний підходи, які дозволили проаналізувати проблему як комплексний управлінський виклик, зокрема в частині інституційного дизайну системи SSA.

Виклад основного матеріалу. В 2010-х роках Управління соціального забезпечення США (далі – Social Security Administration, SSA) – один із ключових федеральних органів, відповідальний за пенсійне та соціальне страхування мільйонів американців – опинилося в стані глибокої операційної кризи. Ключовою проблемою став процес розгляду апеляційних заяв на отримання допомоги по непрацездатності (Disability Insurance). Згідно з офіційними звітами Рахункової палати США (далі - U.S. Government Accountability Office, GAO), кількість нерозглянутих справ на 2017 рік перевищила 1 мільйон [19, 6], а середній час очікування на рішення складав близько 600 днів [16].

Така ситуація мала вкрай негативні соціальні та економічні наслідки. Громадяни, які втратили працездатність, залишалися без засобів до існування на тривалий період, що призводило до зубожіння та погіршення стану здоров'я. Водночас, система була перевантажена, що призводило до високого рівня стресу та плинності кадрів серед фахівців SSA, а також створювало підґрунтя для помилок. Причини кризи були комплексними: старіння населення, ускладнення медичних критеріїв для визначення непрацездатності та застарілі, переважно паперові, процеси обробки справ, кожна з яких могла налічувати сотні й тисячі сторінок медичної документації. Стало очевидним, що екстенсивний шлях (просте збільшення штату) не вирішить проблему. Потрібна була технологічна революція.

У відповідь на кризу SSA ініціювало багаторічний проєкт з модернізації, центральним елементом якого стало створення інтелектуальної системи для обробки заяв. Важливо підкреслити, що розробники свідомо відмовилися від ідеї створення повністю автономного ШІ, який би самостійно приймав рішення «схвалити» чи «відхилити». Така модель була визнана надто ризикованою з етичної точки зору та неприйнятною для суспільства [11]. Натомість була обрана модель «ШІ-асистент» або «система підтримки прийняття рішень». Цей підхід відповідає сучасному етапу розвитку технології, характерною ознакою якого стало «зміщення центру уваги дослідників зі створення автономно функціонуючих систем... до створення людино-машинних систем, інтегрованих в єдине ціле – інтелект людини і здатності обчислювальної машини» [2, с. 179].

Технологічне ядро системи складається з кількох модулів (перелічені деякі):

- Модуль оцифрування та розпізнавання (OCR): Переводить мільйони сторінок паперових медичних записів у машиночитний текст.

- Модуль обробки природної мови (NLP): Це «мозок» системи. Він аналізує текстові масиви, використовуючи методи розпізнавання іменованих сутностей (Named Entity Recognition, NER) для автоматичного виявлення та класифікації ключових елементів: діагнозів, симптомів, результатів аналізів, призначень лікарів, дат проведення процедур.
- Аналітичний модуль на основі машинного навчання (ML): Оброблена інформація порівнюється з тисячами правил та критеріїв, що містяться в нормативній базі SSA. Модель, навчена на даних великої кількості кейсів-справ, оцінює повноту поданої інформації та ймовірність того, що справа відповідає критеріям для призначення виплат.

Процес роботи виглядає наступним чином: після завантаження справи в систему, ШІ-модуль автоматично її аналізує з використанням низки наявних модулів та визначає пріоритетність із внутрішніми метриками, виконуючи умовний скоринг. Після чого спрямовує справу за одним із трьох маршрутів (умовних «кольорових» коридорів):

1. «Зелений коридор»: Справи з високою ймовірністю відповідності критеріям та повною документацією. Вони отримують найвищий пріоритет і негайно передаються фахівцю для фінального, прискореного затвердження.
2. «Жовтий коридор»: Справи, де бракує певних документів або містяться незначні суперечності. Система автоматично формує запит на надання додаткової інформації або позначає конкретні пункти для ретельної уваги фахівця.
3. «Червоний коридор»: Справи з низькою ймовірністю відповідності або серйозними суперечностями в документах. Вони також передаються фахівцю, але з позначкою про необхідність глибокого аналізу.

Впровадження системи призвело до значних позитивних змін. Згідно з внутрішніми звітами SSA, середній час обробки апеляційних справ почав помітно скорочуватися вже з 2018 року і до 2022 року стабілізувався на відмітці 320-340 днів [19, стор.7]. Це дозволило розвантажити чергу та надати допомогу найбільш нужденним категоріям громадян значно швидше.

Окрім прямих кількісних показників, проєкт мав і важливі якісні наслідки. По-перше, він дозволив переорієнтувати роботу фахівців: замість рутинного вичитування тисяч сторінок тексту, вони змогли сфокусуватися на найскладніших, неоднозначних випадках, де потрібен саме людський досвід та емпатія. По-друге, стандартизація процесу аналізу знизила ризик помилок, пов'язаних із людським фактором (наприклад, пропуск важливого діагнозу в об'ємному документі).

Таким чином, можна зробити висновок, що успіх був зумовлений не стільки досконалістю самих алгоритмів, скільки філософією їх застосування. ШІ не замінив людину, а став потужним інструментом у її руках, що дозволило системі публічного управління впоратися з кризою та підвищити свою ефективність. Проте, навіть в таких добре продуманих системах, залишається цілий пласт ризиків, що вимагає глибокого аналізу.

Додатково слід зазначити, що, крім імплементації використання ШІ, для досягнення результатів, SSA було використано також і інші методи, аналіз яких виходить за рамки статті. Саме тому ізольований вплив саме ШІ на кінцеві результат неможливо розрахувати точно, хоча він, вочевидь є дуже суттєвим.

Впровадження алгоритмічних систем у процес прийняття державних рішень неминуче ставить питання про природу та наслідки їхніх помилок. На відміну від традиційних помилок, що допускаються уповноваженим співробітником, помилки ШІ мають потенціал бути систематичними, масштабними та менш очевидними для виявлення. Ефективність, яку обіцяє технологія, може бути швидко нівельована соціальною та фінансовою шкодою від її збоїв, якщо не буде створено надійної системи управління ризиками. Як зазначає Т.С. Яровой, «разом з можливостями, використання штучного інтелекту також відноситься до ризикових аспектів», які включають етичні питання, проблеми конфіденційності даних та відповідальності [4, с. 36]. Аналіз цих ризиків вимагає їх чіткої класифікації, оскільки механізми запобігання та реагування для різних типів помилок будуть докорінно відрізнятися.

В системах, що приймають рішення про надання соціальних благ, доцільно виділити два класичні типи помилок, відомі зі статистики [18], а також третій, «прихований» тип, що стосується системної упередженості.

Помилка I роду (False Negative): «Відмовити тому, хто потребує»

В термінології статистичної перевірки гіпотез, помилка першого роду – це невірне відхилення істинної нуль-гіпотези. У нашому контексті нуль-гіпотеза звучить як «громадянин відповідає критеріям для отримання виплати». Таким чином, помилка першого роду – це неправомірна відмова

особі, що насправді має право на допомогу. З етичної та правозахисної точки зору, це найтяжчий тип помилки, оскільки він безпосередньо порушує право людини на соціальний захист, гарантоване національним законодавством та міжнародними актами.

Наслідки такої помилки є глибоко руйнівними для індивіда та його родини. Вони виходять далеко за межі простої втрати доходу і можуть включати втрату житла, неможливість отримати необхідне лікування, погіршення стану здоров'я через стрес та недоїдання. Для держави така помилка, хоч і не несе прямих фінансових втрат, у довгостроковій перспективі призводить до значно більших витрат через погіршення громадського здоров'я та соціальну маргіналізацію [13].

Враховуючи критичність цього ризику, архітектура системи SSA була розроблена з кількома рівнями захисту:

- Принцип обов'язкового людського контролю (Human-in-the-Loop): Це фундаментальний запобіжник. Жодна рекомендація ШІ про відмову не є остаточною і не призводить до автоматичної дії. Кожна така справа в обов'язковому порядку передається на розгляд кваліфікованому фахівцю. Як зазначають дослідники Адміністративної конференції США, цей принцип перетворює ШІ з «судді» на «асистента-дослідника», який лише готує інформаційну базу для рішення, відповідальність за яке несе людина [11].

- Право на апеляцію як інституційна гарантія: Навіть якщо людський фактор на етапі первинного розгляду призведе до помилкової відмови, у громадянина залишається непорушне право на оскарження. Адміністративна процедура в США передбачає кілька послідовних етапів: 1) перегляд (Reconsideration) іншим фахівцем того ж відомства; 2) слухання в судді з адміністративних справ (Administrative Law Judge), що є квазі-судовим розглядом; 3) розгляд Апеляційною радою; 4) подання позову до федерального суду. Ця багаторівнева система є головною страховкою від того, щоб одна помилка (неважливо, людська чи машинна) стала для людини фатальною.

Помилка II роду (False Positive): «Заплатити шахраю»

Помилка другого роду – це невірне прийняття хибної нуль-гіпотези, тобто схвалення заявки від особи, що насправді не відповідає критеріям або є шахраєм. Наслідки цієї помилки мають іншу природу: вони підривають фінансову стабільність системи соціального забезпечення та руйнують довіру суспільства до держави. Якщо громадяни-платники податків вважають, що система є неефективною і корумпованою, її суспільна легітимність стрімко падає.

Механізми протидії цьому ризику є переважно ретроспективними, тобто спрямованими на виявлення помилки вже після того, як вона сталася:

- Профілактика на етапі схвалення: Хоча ШІ може рекомендувати справу до схвалення, фінальне рішення також приймає людина. Проте, як буде показано нижче, саме тут криється ризик «автоматичної упередженості», коли оператор схильний довіряти «позитивній» рекомендації системи.

- Пост-аудит та інтелектуальний моніторинг: Це основна лінія захисту. Управління SSA використовує окремі аналітичні системи та моделі ШІ, які не працюють з індивідуальними заявками, а аналізують увесь масив вже призначених виплат, шукаючи статистичні аномалії та шахрайські патерни (fraud patterns). Наприклад, система може згенерувати тривожний сигнал (red flag), якщо виявить, що один і той же лікар надав медичні висновки для сотень успішних заявок у різних штатах, або що виплати для десятків бенефіціарів надходять на один банківський рахунок.

- Незалежний нагляд: Офіс Генерального інспектора (OIG) при SSA має повноваження проводити власні розслідування, ініційовані як сигналами від аналітичних систем, так і зверненнями громадян. Згідно з їхнім звітом, використання аналітики даних дозволило виявити та запобігти неправомірним виплатам на сотні мільйонів доларів [15].

- Міжвідомча взаємодія: Дані SSA періодично звіряються з базами даних інших федеральних відомств, наприклад, Податкової служби (IRS). Якщо особа, що отримує допомогу по повній непрацездатності, раптом починає декларувати доходи від повноцінної трудової діяльності, це стає підставою для негайної перевірки.

«Приховані» помилки: Системна упередженість (algorithmic bias)

Найбільш підступний тип помилок не пов'язаний з некоректною роботою коду, а закладений у самій основі машинного навчання – в даних, на яких тренується модель. Якщо історичні дані, що використовувалися для навчання ШІ, відображають існуючу в суспільстві дискримінацію, алгоритм неминуче її засвоїть і почне відтворювати, надаючи цьому процесу ореол об'єктивності та неупередженості [13].

Наприклад, якщо в минулому фахівці-люди з певних соціо-економічних причин (наприклад,

через нижчу якість медичних документів у представників бідніших верств населення) частіше відмовляли певним демографічним групам, модель, навчена на цих рішеннях, може «зробити висновок», що приналежність до цієї групи є фактором ризику, навіть якщо сама змінна «раса» чи «етнічність» буде видалена з датасету. Алгоритм знайде кореляцію через інші, опосередковані змінні (поштовий індекс, тип медичного закладу тощо).

Наслідком стає системна, масштабна та важкодоказова дискримінація, замаскована під об'єктивний технологічний процес. Протидія цьому вимагає комплексних заходів ще на етапі розробки системи:

- Ретельний аудит навчальних даних на предмет прихованих упереджень.
- Використання технік «справедливого машинного навчання» (Fairness-Aware Machine Learning), які спеціально коригують алгоритм для досягнення справедливих результатів для різних демографічних груп.
- Незалежне тестування та сертифікація моделей перед їх впровадженням.

Отже, управління помилками в системах ШІ виходить далеко за межі простого технічного тестування. Воно вимагає побудови комплексної архітектури, що поєднує процедурні гарантії (право на апеляцію), постійний моніторинг, незалежний нагляд та глибоку роботу з даними для запобігання системній дискримінації.

Навіть за умови створення технічно досконалого алгоритму та наявності багаторівневих юридичних процедур оскарження, ефективність та безпечність системи ШІ в публічному управлінні зрештою визначається якістю її взаємодії з кінцевим користувачем — державним службовцем. Практика показує, що саме цей стик «людина-машина» є найбільш вразливою ланкою, де виникають непередбачувані ризики. Ці ризики мають не стільки технічну, скільки когнітивно-психологічну та організаційну природу. Ігнорування цих факторів може призвести до того, що де-юре система буде дорадчою, а де-факто – перетвориться на неконтрольованого автоматизованого диктатора, що ховається за формальним людським затвердженням.

Найбільш дослідженим ризиком у цій сфері є феномен «автоматичної упередженості» – когнітивне викривлення, що полягає у схильності людини надмірно довіряти рекомендаціям, згенерованим автоматизованими системами. Цей феномен, детально описаний у працях з інженерної психології та людських факторів, проявляється у двох основних формах:

1. Помилки дії (errors of commission): Виникають, коли людина-оператор сліпо виконує некоректну рекомендацію системи. Наприклад, фахівець SSA, бачачи, що система позначила шахрайську заявку як «високопріоритетну», може затвердити її без належної перевірки, виходячи з припущення, що «машина розібралася краще».

2. Помилки бездіяльності (errors of omission): Трапляються, коли автоматизована система не виявляє проблему, і людина, покладаючись на мовчання системи, також її не помічає. Наприклад, якщо ШІ не позначив справу як підозрілу, оператор може провести лише поверхневий огляд і не звернути уваги на очевидні суперечності в документах, які він помітив би за умови більш ретельного аналізу [14].

В умовах державної служби, що характеризується високим робочим навантаженням, жорсткими часовими рамками та монотонністю багатьох завдань, ризик виникнення автоматичної упередженості є надзвичайно високим. У службовця з'являється сильна спокуса делегувати когнітивне зусилля машині та перетворити свою роботу на механічне «проклацування» рішень, запропонованих алгоритмом. Це, в свою чергу, повністю нівелює ключовий запобіжник системи – принцип «людини в контурі управління» (human-in-the-loop, HITL). Де-юре відповідальність залишається на людині, але де-факто вона не здійснює осмисленого контролю, а лише легітимізує рішення машини.

Проблема автоматичної упередженості значно поглиблюється, коли система ШІ є «чорною скринькою» (black box). Цей термін використовується для опису моделей (зокрема, складних нейронних мереж), які здатні демонструвати високу точність прогнозів, але внутрішня логіка їхньої роботи є непрозорою та неінтерпретованою для людини. Оператор бачить вхідні дані та кінцевий результат (рекомендацію), але не розуміє, як саме система дійшла такого висновку.

Для сфери публічного управління така непрозорість є неприйнятною з кількох фундаментальних причин:

- Підрив осмисленого нагляду: Якщо фахівець не знає, чому ШІ рекомендував відмовити у виплаті, він не може ефективно перевірити це рішення. Єдине, що він може зробити – це провести повний аналіз справи з нуля, що нівелює саму ідею використання ШІ для економії часу.
- Порушення права на справедливу процедуру: Адміністративна юстиція базується на принципі, що будь-яке державне рішення має бути обґрунтованим. Громадянин має право знати

не лише яке рішення було прийняте щодо нього, але й на яких підставах. Рішення, згенероване «чорною скринькою», за визначенням не може бути обґрунтованим.

- Неможливість вдосконалення системи: Якщо модель починає робити систематичні помилки, розробники не можуть їх ефективно виправити, не розуміючи першопричину аномальної поведінки алгоритму.

Відповіддю на цей виклик стала концепція пояснюваного штучного інтелекту (Explainable AI, XAI) – напрям досліджень, сфокусований на розробці методів, що роблять рішення ШІ зрозумілими для людей [10]. Замість простої відповіді «схвалити/відхилити», система XAI надає додаткову інформацію, наприклад:

- Визначення важливості факторів: Система може підсвітити, які саме факти з медичної справи найбільше вплинули на її рекомендацію (наприклад, *«висновок лікаря-невролога від 15.05.2025»* та «*відсутність результатів МРТ*»).

- Контрфактичні пояснення: Алгоритм може показати, що потрібно було б змінити у заяві для отримання іншого результату (наприклад, *«Якби був наданий висновок про проходження реабілітації, рекомендація змінилася б на «схвалити» з імовірністю 85%»*).

Наявність таких пояснень перетворює взаємодію з ШІ з акту сліпої довіри на осмислений діалог, де людина може критично оцінити аргументи машини та прийняти фінальне, обґрунтоване рішення.

Коли помилка внаслідок роботи ШІ все ж таки трапляється і завдає шкоди, постає одне з найскладніших питань у сфері AI governance: хто несе за це відповідальність? Традиційні правові концепції відповідальності розмиваються у складному ланцюжку взаємодії. Чи винен:

- Розробник/постачальник ПЗ, який створив алгоритм з прихованими упередженнями або недостатнім рівнем надійності?

- Державний орган, який закупив та впровадив цю систему, можливо, не провівши її належного тестування або не забезпечивши адекватну підготовку персоналу?

- Кінцевий користувач-службовець, який, можливо, проявив недбалість і не перевірив рекомендацію системи?

Однозначного підходу до вирішення цієї проблеми в публічному управлінні поки що не визначено. Публічне управління лише починає формувати підходи. Одним із таких підходів є концепція «осмисленого людського контролю» (meaningful human control) як критерію для покладання відповідальності. Згідно з цією концепцією, відповідальність може бути покладена на людину-оператора лише в тому випадку, якщо вона мала реальну, а не ілюзорну, можливість вплинути на рішення. А така можливість виникає лише за дотримання двох умов: 1) система є прозорою та пояснюваною (XAI); 2) оператор має достатньо часу, ресурсів та кваліфікації для критичної оцінки її рекомендацій [12].

В іншому випадку відповідальність має покладатися на організацію в цілому – тобто на державний орган, який зобов'язаний вибудувати всю систему (технологія + процедури + навчання) таким чином, щоб мінімізувати ризики.

Висновки. Таким чином, розглянувши існуючі проблеми та ризики впровадження ШІ у публічне управління, можемо зробити висновок, що безпечне та ефективне впровадження штучного інтелекту в публічне управління є не стільки технологічним, скільки комплексним інституційним завданням. Досвід Управління соціального забезпечення США наочно демонструє, що найбільш дієвою моделлю є «ШІ-асистент», який не замінює, а посилює людину, залишаючи за нею остаточне рішення. Такий підхід, що базується на принципі «людини в контурі», є ключовим запобіжником від неконтрольованих алгоритмічних помилок, які мають потенціал бути системними, масштабними та менш очевидними, ніж традиційні людські хиби.

Аналіз показав, що найбільш вразливою ланкою в усьому ланцюжку прийняття рішень є взаємодія на стику «людина-алгоритм». Такі когнітивні пастки, як «автоматична упередженість», та технічні проблеми, як-от непрозорість «чорних скриньок», здатні нівелювати будь-які формальні процедури контролю. Отже, управління ризиками в системах ШІ вимагає побудови багаторівневої архітектури захисту, що поєднує інституційні гарантії (право на апеляцію), технічну прозорість (пояснюваний ШІ) та глибоку роботу з психологією прийняття рішень, щоб перетворити ШІ на надійного та контрольованого партнера, а не на непередбачуваного «цифрового левіафана» та може стати перспективними напрямками подальших досліджень.

Список літератури:

1. Подворнюк О. О., Поліщук Н. В. Перспективи впровадження електронного документообігу в публічному управлінні. *Вчені записки ТНУ імені В.І. Вернадського. Серія: Державне управління*. 2021. Том 32 (71), № 5. С. 19–23.
2. Воронов О., Остапенко О., Яценко В. Вплив штучного інтелекту на прийняття управлінських рішень у публічному управлінні. *Теоретичні та прикладні питання державотворення.* 2024. Вип. 32. С. 177–187.
3. Оболенський О., Косицька В., Рвач А. Штучний інтелект у публічному управлінні: вимоги, проблеми та ризики. *Вісник Київського національного економічного університету імені Вадима Гетьмана.* 2023. №33. С. 121–137.
4. Яровой Т.С. Можливості та ризики використання штучного інтелекту в публічному управлінні. Науковий журнал *«ECONOMIC SYNERGY»*. 2023. Вип. 2 (8). С. 36–47.
5. Glaze K, Ho D. E. (n.d.). The Social Security Administration and AI Governance. Stanford University. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3935950
6. Shorey S. AI and Government Workers: Use Cases in Public Administration. Roosevelt Institute, 2025. URL: https://rooseveltinstitute.org/publications/ai-and-government-workers
7. Alon-Barkat S., Busuioc M. Human-AI Interactions in Public Sector Decision-Making. URL: https://arxiv.org/abs/2103.02381
8. Bovens M., Zouridis S. From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review.* 2002. № 62(2). 174–184. https://doi.org/10.1111/0033-3352.00168
9. European Commission. 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. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
10. Gunning D., Stefik M., Choi J., Miller T., Stumpf S., & Yang, G.-Z. XAI—Explainable artificial intelligence. *Science Robotics.* 2019. № 4(37). URL: https://www.science.org/doi/10.1126/scirobotics.aay7120
11. King J., Coglianese C. Algorithmic Tools in Government. *Report for the Administrative Conference of the United States.* URL: https://www.acus.gov/report/algorithmic-tools-government
12. Margetts H., Dunleavy P. The 'Digital State' Transformed: How the Internet is Changing Public Services. In The Oxford Handbook of Internet Studies. Oxford University Press, 2013. URL: https://academic.oup.com/book/27333/chapter/197177864
13. O'Neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, 2016. URL: https://weaponsofmathdestructionbook.com/
14. Parasuraman R., Riley V. Humans and automation: Use, misuse, disuse, abuse. *Human Factors.* 1997. № 39(2). P. 230–253. URL: https://journals.sagepub.com/doi/10.1518/001872097778543821
15. SSA Office of the Inspector General. Use of Electronic Data to Address Improper Payments. Audit Reports, 2022. URL: https://oig.ssa.gov/audits-and-investigations/audit-reports/
16. Social Security Disability: Better Timeliness Metrics Needed to Assess Transfers of Appeals Work (GAO-18-501) (2018). URL: https://www.gao.gov/assets/gao-18-501.pdf
17. Міністерство цифрової трансформації України. Біла книга «Регулювання штучного інтелекту: український контекст та рекомендації» (2024). URL: https://thedigital.gov.ua/storage/uploads/files/page/community/docs/%D0%A0%D0%B5%D0%B3%D1%83%D0%BB%D1%8E%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F%20%D0%A8%D0%86.pdf
18. Neyman J., Pearson E. S. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society.* 1933. № 231. P. 289–337. URL: https://doi.org/10.1098/rsta.1933.0009
19. The Social Security Administration's Hearings Backlog and Average Processing Times. URL: https://oig.ssa.gov/assets/uploads/a-05-22-51159r.pdf

Shyshkin S. O. EXPERIENCE OF IMPLEMENTING AI IN THE US SOCIAL SECURITY SYSTEM AS A RISK MINIMIZATION TOOL

The article analyzes the nature and consequences of errors that arise when applying artificial intelligence (AI) systems in public administration. Using a detailed analysis of the modernization of the US Social Security Administration (SSA) as an example, the article examines the successful model of the «AI assistant,» based on the principle of «human in the loop» as a key tool for risk minimization.

The article proposes a comprehensive classification of potential errors, namely: first-order errors («denial to those in need»), which pose a direct threat to human rights; second-order errors («payments to fraudsters»), which undermine financial stability and trust in state institutions; «hidden» errors of systemic bias (algorithmic bias), which arise from historical data and are capable of reproducing social inequality under the guise of

objectivity. Particular attention is paid to the most vulnerable link in the system — interaction at the «human-algorithm» interface. Cognitive traps such as «automation bias» and the «black box» problem are analyzed, which can nullify any formal precautions and turn human control into a ritual act. It has been proven that the effective and safe implementation of AI is not so much a technological challenge as a complex institutional one, requiring the construction of a multi-level protection architecture that combines procedural guarantees, technical transparency (XAI), and in-depth work on the psychology of decision-making. This approach will allow turning AI into a reliable partner of the state, preventing its transformation into an uncontrolled "digital leviathan" that threatens democratic institutions.

Key words: *public administration, decision-making, artificial intelligence, automatic bias phenomenon, explainable AI, human control, AI regulation, administrative services.*

Дата надходження статті: 05.09.2025
Дата прийняття статті: 18.09.2025
Опубліковано: 29.12.2025