Лабораторія
Цифрової
Безпеки

ВИКОРИСТАННЯ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ ВІДПОВІДНО ДО ПРАВ ЛЮДИНИ

ПОСІБНИК ДЛЯ ГРОМАДЯНСЬКОГО СУСПІЛЬСТВА

Тетяна Авдєєва,
Олександр Батхін

2024

Цей посібник окреслює ключові виклики та ризики для прав людини, з якими стикаються громадянське суспільство, журналісти та правозахисники в усьому світі при застосуванні систем штучного інтелекту. У ньому наводяться конкретні приклади загроз безпеці персональних даних, рівності, свободі вираження поглядів та іншим пов'язаним з ними свободам, а також детально розглядаються способи уникнення або зменшення таких ризиків. Посібник також містить керівництво для відповідального вибору систем ШІ, яке дозволить читачам швидко перевірити, чи відповідає їхня система штучного інтелекту основним стандартам у сфері прав людини, і переконатися, що вибір технології є усвідомленим, обачним і відповідальним.

Лабораторія цифрової безпеки (Цифролаба) - українська неурядова організація, що допомагає незалежним медіа, журналістам, активістам та громадянському суспільству посилювати власну цифрову безпеку. Цифролаба також працює над впровадженням стандартів прав людини в цифровій сфері через напрацювання аналітичних матеріалів, долучення до законотворчих процесів та адвокаційних кампаній в Україні та світі.

Наші контакти:
https://dslua.org/
dslua@dslua.org
ФБ: https://www.facebook.com/dslua
X: @DSLab_Ukraine

ЗМІСТ

ВСТУП

Зростаюча популярність систем штучного інтелекту (ШІ) серед організацій громадянського суспільства (ОГС) надає різноманітні переваги, а також викликає численні занепокоєння щодо впливу на права людини та дотримання відповідних стандартів. Дуже важливо запобігти небезпечному та нерегульованому використанню систем ШІ, яке може призвести до різних негативних наслідків для громадянського суспільства.

Чому громадянське суспільство використовує інструменти ШІ? Швидка цифровізація надала всім користувачам різноманітні автоматизовані інструменти, починаючи від зовнішніх інструментів, що впливають на їхні права, і закінчуючи системами, які вони свідомо використовують для підвищення продуктивності. Громадянське суспільство не є винятком, і бажає отримати переваги від систем ШІ, такі як доступ до швидкого створення контенту, ефективні інструменти дослідження, ефективна ідентифікація фейкових новин і багато інших переваг. У той же час воно є більш вразливим до ризиків і небезпек, пов'язаних із зловживанням і безвідповідальним використанням ШІ, що відкриває шляхи для безлічі нових ризиків і проблем. Цей посібник розглядає системи ШІ в їхньому найширшому значенні, представленому ОЕСР.

> **Система ШІ** - машинна система, яка для прямих або опосередкованих цілей на основі отриманих вхідних даних генерує результати, такі як передбачення, контент, рекомендації або рішення, які можуть впливати на фізичне або віртуальне середовище.

Які виклики несе в собі безвідповідальне використання ШІ? Безвідповідальне використання систем ШІ може сприяти поширенню дезінформації, витоку даних, порушенням прав інтелектуальної власності, алгоритмічній упередженості та іншим небезпечним результатам. Громадянське суспільство може підсвідомо посилювати такі ризики, якщо їм не вистачає належного людського нагляду за використанням систем ШІ, базових знань про технічну сторону та усвідомлення ключових проблем на ринку ШІ. Різноманітний набір прикладів, коли неконтрольоване або погано кероване використання ШІ призвело до порушень прав людини, включає:

- Витік даних про те, що ШІ-генератор відео Runway, який фінансується Google, навчався на краденому контенті з YouTube і піратських фільмах;
- Діпфейк про те, як прозахідна президентка Молдови підтримує політичну партію, що поширює проросійські наративи;
- ChatGPT зазнав витоку чутливих даних користувачів, ймовірно, після злому;

- Випадкове розкриття працівниками Samsung комерційної таємниці через ChatGPT;
- Колишній інженер Google заарештований за крадіжку секретів систем ШІ Google для китайських компаній;
- Система рекрутингу на основі ШІ iTutor Group, яка відхиляє кандидатів, дискримінуючи за віком;
- Алгоритми охорони здоров'я які використовуються лікарнями та страховими компаніями, не позначають людей з нетиповим для регіону кольором шкіри як пацієнтів.

До чого варто прагнути ОГС, які вдаються до використання інструментів на основі ШІ? ОГС, які вирішують використовувати системи ШІ, повинні уникати або пом'якшувати всі ризики, пов'язані з такими інструментами, шляхом впровадження стандартів відповідального та законного використання ШІ. Такі вимоги охоплюють, наприклад, забезпечення дотримання міжнародних стандартів прав людини та відповідних національних нормативних актів, безпечне використання систем ШІ, їх нейтральність і прозорість, а також встановлення належного людського нагляду, здійснення оцінки впливу на права людини та впровадження процедур оцінки ризиків.

Метою Посібника є запобігання або пом'якшення ризиків, пов'язаних із використанням ОГС систем ШІ. Надаючи рекомендації щодо безпечного використання як зовнішніх систем ШІ (наприклад, ChatGPT, DALL-E, Midjourney), так і внутрішніх інструментів на основі ШІ (тобто розроблених, замовлених або налаштованих ОГС), набір інструментів спрямований на забезпечення дотримання відповідних нормативних стандартів і вимог до прав людини, які відображають сучасні стандарти, встановлені Рамковою конвенцією про штучний інтелект і права людини, демократію та верховенство права, а також Законом ЄС про штучний інтелект. Рекомендації також надають керівництво, що дозволяє здійснювати відповідальний вибір систем ШІ.

РОЗДІЛ 1. ОЦІНКА ВПЛИВУ НА ПРАВА ЛЮДИНИ ТА ОЦІНКА РИЗИКІВ

Оцінка впливу на права людини (ОВПЛ) — це процес виявлення, аналізу та усунення несприятливих наслідків систем ШІ на реалізацію прав людини будь-якими зацікавленими сторонами (включно з користувачами систем ШІ, членами інших ОГС, членами вразливих спільнот тощо). ОВПЛ та загальна оцінка ризиків (ОР) є важливими для будь-якої ОГС, яка бажає використовувати переваги систем ШІ та запобігати небезпекам, пов'язаним з ними. У цьому контексті особливо важливо уникати заміни ОВПЛ простим аналізом ризиків для організаційної моделі, кібербезпеки, фінансової безпеки або перевіркою дотримання національного законодавства.

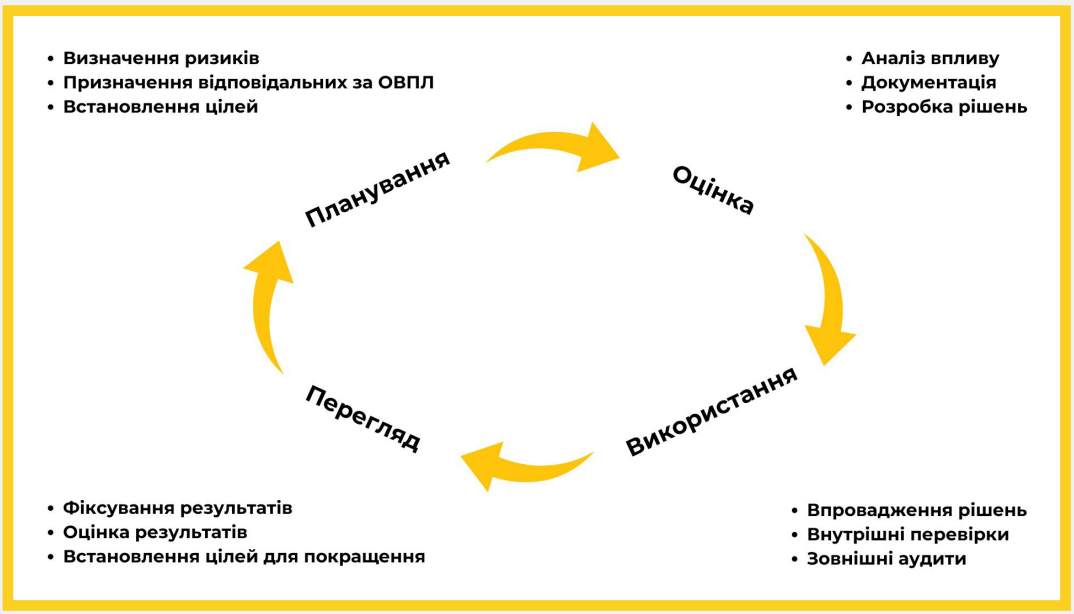
Різниця між ОВПЛ та ОР:	
ОВПЛ	**ОР**
Проводиться на всіх етапах життєвого циклу системи ШІ	
Включає внутрішні перевірки та зовнішні аудити	
Стосується як позитивного, так і негативного впливу системи ШІ на права людини	Стосується недоліків систем ШІ для бізнес-стратегії, організаційної моделі, кібербезпеки та інших аспектів діяльності
Стосується практики використання конкретної системи ШІ	Стосується загального режиму роботи організації
Проводиться регулярно, в тому числі після оновлення системи ШІ, скарг користувачів, нормативних змін тощо	Проводиться регулярно протягом усієї діяльності ОГС (не прив'язана до використання систем ШІ)

Відповідно, обидві процедури мають бути включені в практику ОГС із запровадженням відповідних методологій і протоколів для забезпечення старанного та ретельного аналізу ризиків і наслідків. Це також передбачає перегляд системи підзвітності та відповідальності всередині організації.

Рекомендації. Для запобігання та пом'якшення ризиків, пов'язаних із системами ШІ, ОГС має впроваджувати ОВПЛ та ОР на всіх етапах життєвого циклу системи ШІ, тобто дослідження, розробки, розгортання, використання, завершення використання, демонтажу та припинення. Будь-яка ОГС, яка розробляє або використовує системи ШІ, повинна проводити перевірки відповідності систем ШІ застосовним нормам щодо захисту персональних даних (ПД), свободи вираження поглядів, рівності, прав інтелектуальної власності тощо. Зокрема, вони повинні вживати наступних кроків:

- **Обирати системи ШІ відповідально.** Перш ніж включати системи ШІ в роботу ОГС, краще уникнути очевидних ризиків, відповідально обираючи такі інструменти. Це може включати утримання від використання певних систем ШІ, наприклад, тих, які явно порушують права людини. Детальну інструкцію щодо відповідального вибору інструментів ШІ можна знайти в **Розділі 7 цього Посібника (сторінка 25)**;
- **Розробляти ефективні методології.** Для того, щоб ОГС належним чином проводили ОВПЛ та ОР, потрібні внутрішні методології для обох оцінок. Такі методології мають бути сучасними, доступними для команди та визначати алгоритм того, як оцінюються ризики та наслідки, хто відповідає за таку оцінку та як створюються рішення;
- **Призначати відповідальних осіб.** Для проведення ОВПЛ та ОР ОГС має призначити відповідальних осіб, які координуватимуть, здійснюватимуть нагляд та проводитимуть внутрішні аудити та розроблятимуть стратегії зменшення ризиків, а також призначених членів команди, які будуть впроваджувати результати ОВПЛ та ОР, пом'якшуючи негативні наслідки та ризики для ОГС;
- **Проводити в організації ретельну та своєчасну ОР.** ОГС повинні регулярно оцінювати ризики, пов'язані з їхньою діяльністю, наприклад фінансові, юридичні, безпекові, ризики продуктивності тощо, і розробляти рішення, які усувають або пом'якшують такі загрози. Безперервний процес ОР має включати щоквартальні перегляди ризиків і планів управління ризиками, а також щорічні перегляди політик, підходів і критеріїв оцінки ризиків. Інтенсивність таких перевірок залежить від рівнів ризику, а також слід проводити додаткові перевірки після інцидентів, зміни політики в ОГС або нововиявлених небезпек;
- **Проводити ретельну та своєчасну ОВПЛ для всіх систем ШІ.** ОГС має проводити ОВПЛ, щоб переглядати, зважувати та збалансовувати позитивні та негативні впливи систем ШІ та знаходити способи усунення або пом'якшення загроз. Варто пересвідчитися, що внутрішні та зовнішні перевірки проводяться якомога раніше відносно початку нового проекту, наприклад, з етапу дизайну або розробки системи ШІ. Регулярні перевірки впливу систем ШІ на права людини повинні проводитися принаймні раз на рік з додатковими перевірками в момент оновлення або нормативних змін, а також після будь-якої відповідної скарги;
- **Проводити внутрішні і зовнішні аудити.** ОГС має регулярно організовувати внутрішні аудити та дозволяти здійснювати зовнішні аудити для здійснення ОВПЛ та ОР зовнішніми незалежними професіоналами та експертами всередині організації, особливо якщо їхній власний досвід у певній темі не дозволяє їм суттєво визначити всі ризики та наслідки;

- **Відповідально обирати зовнішніх аудиторів.** Залежно від сфери проведення ОР та ОВПЛ, ОГС повинні ретельно обирати компанії для зовнішніх аудитів, а також переглядати їх досвід, репутацію та методології ОР чи ОВПЛ відповідно до потреб, цінностей та цілей ОГС;
- **Належно впроваджувати результати ОВПЛ та ОР.** Після проведення ОВПЛ та ОР, ОГС має впроваджувати всі рішення, розроблені за допомогою цих процесів, усуваючи або пом'якшуючи ризики та негативні впливи на права людини. Впроваджуючи рішення ОВПЛ та ОР, ОГС має завжди покращувати свої зусилля для встановлення кращих практик для себе. Повний цикл ОВПЛ та ОР повинен відповідати цьому алгоритму:



- **Своєчасно повідомляти розробників або постачальників зовнішніх систем ШІ про виявлені ризики та несприятливі наслідки.** Якщо під час ОВПЛ ОГС виявить ризики, несприятливий вплив або інші проблеми з системами ШІ, вона повинна негайно повідомити розробника або постачальника таких систем про ці проблеми, щоб розробник або постачальник покращив ситуацію. Якщо реакції не буде, є розумні підстави розглянути альтернативні системи ШІ;
- **Регулярно переглядати методології.** Організація повинна встановити власну методологію перевірки для належного проведення ОВПЛ. Наприклад, Данський інститут прав людини пропонує комплексний підхід до методології ОВПЛ, який складається з п'яти основних етапів:

Планування та визначення обсягу	Спеціалісти-практики з ОВПЛ повинні визначити відповідних зацікавлених сторін для консультацій протягом усього процесу ОВПЛ. Крім того, також можуть проводитися попередні інтерв'ю із зацікавленими сторонами.

Збір даних і розробка бази	Спеціалісти-практики з ОВПЛ виїжджають на місця, щоб дослідити рівень дотримання прав людини працівниками, членами громади та іншими відповідними правовласниками. На цьому етапі особливе значення приділяється роботі на місцях, інтерв'ю та іншим типам активного залучення зацікавлених сторін.
Аналіз впливу	Спеціалісти-практики з ОВПЛ повинні проаналізувати зібрані дані, щоб виявити будь-які впливи на права людини та оцінити їх рівень. Цей етап охоплює оцінку міжнародних стандартів і принципів прав людини, порівняльні проекти тощо відповідно до результатів залучення зацікавлених сторін.
Пом'якшення впливу та управління впливом	ОГС, спеціалісти-практики з ОВПЛ та зацікавлені сторони повинні об'єднати свої зусилля для створення плану запобігання та усунення наслідків для прав людини, віддаючи пріоритет найсерйознішим із них.
Звітність та оцінка	Спеціалісти-практики з ОВПЛ надають детальний звіт про ОВПЛ, який є доступним для правовласників, носіїв обов'язків та інших відповідних сторін.

- **Проводити перевірки відповідності.** Потрібно забезпечити дотримання прав людини в системах ШІ та їх відповідність чинним національним і міжнародним нормам. Наприклад, Закон ЄС про штучний інтелект встановлює вимоги щодо прозорості, маркування, людського нагляду та оцінки ризиків для тих, хто використовує системи ШІ. Крім того, Загальний регламент захисту даних (GDPR) визначає вимогу щодо чіткої згоди на обробку ПД. Крім того, життєво важливо дотримуватися сучасних уніфікованих підходів до ОВПЛ та ОР для систем ШІ, таких як HUDERIA, що спрямована на надання чітких, конкретних та об'єктивних критеріїв оцінки та пом'якшення негативного впливу на права людини, демократію та верховенство права.

РОЗДІЛ 2. ЗАХИСТ ПЕРСОНАЛЬНИХ ДАНИХ ТА ЦИФРОВА БЕЗПЕКА

Обробка даних відіграє вирішальну роль у впровадженні інструментів ШІ в роботу громадянського суспільства. Однак виникають численні ризики як на правовому, так і на технічному рівнях, що створює потенційні негативні впливи на права людини, цифрову безпеку та загальну ефективність роботи ОГС. Таким чином, Лабораторія цифрової безпеки вважає відповідальне управління даними необхідною передумовою для сумісного з правами людини використання інструментів, керованих ШІ.

Захист персональних даних. Більшість систем ШІ збирають, обробляють і зберігають ПД, створюючи різні загрози конфіденційності користувачів, які можуть включати розкриття чутливих даних. Відповідно до індексу AIGS, близько 40% країн світу також активно покладаються на спостереження на основі ШІ. Крім того, зростання популярності технологій розпізнавання обличчя створює нові наслідки для сфери приватного життя. Як приклад порушень права на недоторканість приватного життя, пов'язаних з ШІ, у 2023 році Clearview AI зіткнувся з судовим позовом через масовий скрейпінг ПД із профілів мільйонів користувачів у соціальних мережах без їхньої згоди. Ці технології створюють ризики порушення права на недоторканість приватного життя користувачів, прихованого збору та зберігання даних, зловмисного використання ПД тощо.

Рекомендації. Щоб запобігти порушенням прав інших людей, важливо збирати будь-які дані щодо них лише прозоро та за згодою. Будь-яка ОГС, яка використовує системи ШІ, що працюють з ПД, повинна гарантувати, що обробка даних є консенсусною, прозорою та законною. Ось деякі з кроків, які ОГС повинні вжити, щоб забезпечити захист приватного життя людей:

- **Уникати інструментів на основі ШІ, які вдаються до явно незаконних практик. Слід** уникати систем ШІ, які відомі незаконним збором, зберіганням, обробкою ПД та навчанням на ПД. Крім того, не слід використовувати системи ШІ, які не мають чітких гарантій у своїй політиці конфіденційності або які вдаються до скрепінгу даних, неавторизованої передачі даних користувачів, прихованого збору метаданих тощо;
- **Створити та підтримувати актуальність політик захисту ПД.** ОГС має встановити чітку та прозору політику щодо захисту ПД, доступну як команді, так і громадськості. Така політика захисту ПД повинна охоплювати:
  - Основні принципи захисту даних;
  - Стратегії захисту даних організації, включаючи окремих осіб, відділи, пристрої та IT-середовища;

- Права суб'єкта ПД;
- Правила передачі ПД третім особам;
- Положення щодо відповідності захисту даних законодавству;
- Розподілені ролі та обов'язки, включаючи зберігачів даних і ролі, які відповідають за координацію діяльності із захисту даних;

- **Законно збирати та використовувати ПД у системах ШІ.** Будь-яка ОГС, яка використовує системи ШІ, що збирають ПД, повинна переконатися, що ці системи ШІ мають дійсну правову основу, таку як згода, контракт або суспільний інтерес, для збору та використання ПД. Крім того, системи ШІ повинні надавати користувачам чітку та доступну інформацію про те, як їхні ПД збираються, використовуються та передаються, і які їхні права щодо цих даних;
- **Видаляти дані вчасно та правильно.** Використовуючи будь-які ПД, ОГС має переконатися, що коли мета обробки даних досягнута, ці дані будуть повністю видалені та не відображатимуться знову через систему ШІ; частини даних, які може знадобитися зберегти для звітності або для юридичних цілей, повинні бути належним чином знеособлені (за можливості) та безпечно зберігатися;
- **Законно передавати ПД третім сторонам (за необхідності).** Передаючи ПД третім сторонам, ОГС повинні дотримуватися відповідних вимог GDPR і політики безпеки. Зокрема, ОГС повинні укладати угоди про захист даних із третіми сторонами, які гарантують дотримання ними стандартів захисту даних, а також вести облік процесів передачі даних;
- **Забезпечувати права суб'єктів даних.** Життєво важливо захистити права осіб, чиї дані будь-яким чином обробляються системою ШІ. Зокрема, такі особи повинні:
  - мати доступ до своїх ПД, оброблених системами ШІ;
  - мати змогу заперечувати проти обробки їхніх ПД;
  - мати можливість змінити застарілі чи неточні дані;
  - мати право вимагати видалення даних;
- **Забезпечувати захист третіх осіб.** Якщо ОГС збирає дані за допомогою стеження, розпізнавання обличчя, OSINT-технологій тощо (наприклад, у рамках антикорупційних розслідувань), життєво важливо захистити третіх осіб від викриття та непотрібного збору їхніх ПД. Наприклад, під час використання зображень, отриманих за допомогою технологій стеження, усі особи, окрім об'єкта розслідування, мають бути знеособлені шляхом розмиття їхнього обличчя чи інших ідентифікаційних ознак;
- **Уникати використання ПД у загальнодоступних системах ШІ**. Оскільки системи ШІ часто зберігають і повторно обробляють дані, особливо загальнодоступні системи, життєво важливо переконатися, що працівники ОГС уникають використання ПД у підказках або під час навчання системи ШІ;

- **Регулювати доступ до ПД в команді ОГС**. Для забезпечення захисту ПД життєво важливо обмежити доступ до ПД всередині команди та дозволити його лише призначеним відповідальним особам;
- **Належним чином та своєчасно повідомляти про інциденти з даними.** У разі будь-яких інцидентів з даними особа, яка зіткнулася з інцидентом, повинна повідомити про це відповідальних осіб у команді, а також зацікавлених третіх осіб або розробників і постачальників систем ШІ (залежно від характеру інциденту). Окрім повідомлення команди про інцидент, будь-який співробітник, який стикається з інцидентами з даними, повинен вжити відповідних заходів безпеки, про які йдеться нижче.

Цифрова безпека. Незважаючи на зростаючу популярність, ШІ може викликати серйозні проблеми з безпекою для ОГС, особливо щодо захисту даних. Згідно з дослідженням кібербезпеки ШІ, проведеним Департаментом науки, інновацій та технологій Великої Британії, 68% опитаних компаній активно використовують моделі ШІ для свого бізнесу. У той же час 81% зіткнулися з порушенням безпеки свого ШІ або не в змозі виявити вразливість. Використання систем ШІ без належних заходів безпеки може призвести до:

- **Отруєння моделі:** атака на модель ШІ, яка вводить шкідливі дані в навчальну базу даних, змушуючи генеративні моделі ШІ видавати менш точні результати. Іноді отруєння моделей може призвести до безпрецедентної упередженості, дискримінації або дезінформації у контенті, створеному отруєними моделями ШІ;
    - Підмножина атак із отруєнням моделі є навмисною: атаки, які змінюють навчальні дані, щоб схилити модель до певного результату;
- **Змагальні приклади:** модифіковані версії законних вхідних даних, які створені, щоб обдурити модель;
- **Атаки ухилення:** атаки, які обходять системи безпеки, змінюючи вхідні дані, щоб уникнути виявлення або класифікації моделлю;
- **Крадіжка моделей:** атаки з вилученням параметрів або архітектури навченої моделі для створення копії моделі;
- **Витік даних**: інцидент, який призводить до зливу або викрадення інформації з організації. Часто викрадені дані можуть містити чутливу інформацію (наприклад, біометричну інформацію, комерційну чи банківську таємницю або дані, що стосуються питань національної безпеки).

Ризики, які становлять для ОГС вразливі місця в безпеці моделей ШІ, збільшують потребу у встановленні відповідних гарантій безпеки, впровадженні додаткових заходів і розробці методів уникнення або пом'якшення ризиків, пов'язаних із безпекою.

Рекомендації. Щоб запобігти та зменшити безпекові ризики, ОГС повинні впроваджувати процедурні заходи на організаційному рівні, дотримуватися принципів управління даними та чітко розрізняти, які взаємодії з інструментами, керованими ШІ, створюють додаткові ризики безпеці. Зокрема, ОГС повинні взяти на себе зобов'язання щодо наступних кроків:

- **Проводити тренінги з цифрової безпеки (ТЦБ) для команд ОГС.** Організація повинна навчати весь свій персонал основним принципам цифрової безпеки, пояснювати та навчати протоколам безпеки, а також проводити оцінювання, щоб гарантувати, що 100% персоналу захищено від кіберзагроз і навчено правилам кібербезпеки та цифрової гігієни. Теми ТЦБ, серед іншого, можуть охоплювати:
  - Приватність даних;
  - Захист пароля та автентифікація;
  - Тіньові IT-ресурси;
  - Неавторизоване програмне забезпечення та шкідливі програми;
  - Безпека електронної пошти та захист від фішингу;
  - Політика кібербезпеки ОГС та протоколи надзвичайних ситуацій;
  - Інші питання, що стосуються діяльності ОГС.
- **Перевіряти походження інструментів ШІ**. Варто відмовитись від використання систем ШІ, які походять з компаній чи країн з високим індексом порушень прав людини або країн, у яких немає достатніх даних щодо дотримання ними прав людини, особливо тих, які сумно відомі своєю практикою стеження та переслідування. (наприклад, Росія, Китай, Іран, Північна Корея тощо). Докладну інструкцію можна знайти **в Розділі 7 цього Посібнику (сторінка 26)**;
- **Розробляти політики і протоколи безпеки**. ОГС повинна розробити безпекову політику, яка, зокрема, охоплює правила реагування на надзвичайні ситуації. У правилах зазначаються особи, відповідальні за підтримку безпеки даних, а також механізми захисту чутливих даних. Правила реагування на надзвичайні ситуації мають відповідати політикам безпеки даних ОГС, які б разом охоплювали всі ризики та реагування на них;
- **Персоналізувати налаштування інструментів, керованих ШІ**. Варто скасувати дозвіл на використання ПД для подальшого навчання системи, заборонити доступ до файлів на пристрої, відмовитись від хмарного сховища даних і налаштувати корпоративний доступ залежно від посад та відповідальності осіб у команді ОГС. (наприклад, ChatGPT дозволяє користувачам обирати, які попередні введення даних можна використовувати для подальшого навчання системи);

- **Переконатись, що навчальні та тестові набори даних є безпечними та надійними.** ОГС повинні обов'язково перевіряти надійність навчальних та тестових наборів даних, гарантувати, що треті сторони не мають несанкціонованого доступу до таких наборів даних і не можуть неконтрольовано змінювати їхню суть для подальшого впливу на модель ШІ;
- **Прагнути мати окремі персональні та професійні пристрої та профілі**. Якщо можливо, варто утримувати окремі пристрої та профілі додатків для діяльності, пов'язаної з ОГС, і для особистої діяльності, і намагатися уникати їх взаємозамінного використання для тих самих цілей. Персональні пристрої або профілі не повинні містити жодних даних, що стосуються діяльності ОГС, особливо конфіденційної інформації. Якщо підтримувати окремі пристрої та профілі неможливо, обов'язково слід застосовувати посилені заходи безпеки до персональних пристроїв і профілів, які використовуються для роботи, зокрема шифрування пристрою та інші розширені механізми захисту. Потрібно забезпечити їх своєчасне технічне обслуговування власною командою безпеки або довіреними третіми сторонами;
- **Забезпечити захист від фішингу та розробити механізми реагування на інциденти**. Правило номер 1 — ніколи не надавати персональні чи інші конфіденційні дані у відповідь на небажаний запит. Слід уникати натискання посилань, файлів або вкладень, якщо вони викликають підозру. Також варто захистити свої облікові записи, встановивши надійну багатофакторну автентифікацію. ОГС має заохочувати членів команди своєчасно повідомляти про будь-які підозрілі повідомлення та можливі інциденти;
- **Дотримуватись принципу мінімізації даних**. Слід звести до мінімуму використання чутливих даних під час використання систем ШІ, які цього потребують (наприклад, технології розпізнавання обличчя на основі ШІ, біометрія відбитків пальців тощо). Використовуючи такі технології, ОГС повинні застосовувати необхідний мінімум даних, отриманих і використаних на основі консенсусу, які не створюють загрози, якщо зберігаються або виявляються за допомогою системи ШІ;
- **Уникати використання чутливих даних**. Не можна використовувати чутливі дані під час створення вхідних даних для відкритих систем ШІ, які зберігають надані дані. Використання чутливих даних із загальнодоступними моделями ШІ, такими як ChatGPT, PerplexityAI, Gemini та іншими, може призвести до виявлення цих даних звичайними користувачами або доступу зловмисників через несанкціонований доступ та витік даних. Тому краще не використовувати чутливі дані з будь-якою онлайн-системою ШІ, якій ОГС не може повністю довіряти, щоб захистити їх;

- **Запровадити методи знеособлення та шифрування**. Варто анонімізувати дані, замінивши будь-які ідентифікаційні дані випадково згенерованими. В іншому випадку, коли потрібно використовувати ПД, шифрування захищає їх шляхом шифрування ідентифікованих даних за допомогою надійних протоколів, які розшифровуються пізніше. Важливо підкреслити, що на відміну від знеособлення, шифрування не видаляє конфіденційні дані, оскільки їх можна розшифрувати, тому його слід використовувати обережно, а механізм дешифрування, особливо використовуваний ключовий матеріал, має бути захищений;
- **Врегулювати використання віртуальних помічників.** Використання віртуальних помічників викликає занепокоєння щодо безпеки, оскільки доки програму встановлено на пристрої з мікрофоном, ця програма завжди створюватиме ризик фонового прослуховування та подальшої передачі конфіденційної інформації. Тому слід уникати віртуальних помічників на звичайних персональних або професійних пристроях і використовувати їх поблизу конфіденційної інформації. Якщо це неможливо, варто обмежити їх дозволи «лише під час використання» або застосовувати інші обмежувальні налаштування, доступні на відповідній платформі.

РОЗДІЛ 3. ЗАПОБІГАННЯ АЛГОРИТМІЧНІЙ УПЕРЕДЖЕНОСТІ ТА ДИСКРИМІНАЦІЇ

Однією з очевидних проблем систем ШІ є упередженість, яка може виникнути в таких системах. Через отруєння моделі, недостатні навчальні дані або просто невиразний характер самого ШІ він не може автоматично фільтрувати шовінізм і стереотипи зі своїх результатів. Упередженість може вплинути не лише на конкретні види використання, але й на основні послуги, які надають моделі ШІ. Щоб використовувати результати моделей ШІ, потрібен додатковий рівень фільтрації проти упередженості, яку вони можуть містити. Крім того, важливо не тільки уникати упередженості в результатах моделей ШІ, але й взаємодіяти з цими моделями таким чином, щоб забезпечити з часом нижчі рівні дискримінації.

Рекомендації. Життєво важливо оцінити інформацію, яку створює модель ШІ, і ще раз перевірити її, щоб переконатися у відсутності упередженості, а також переконатися, що взаємодія користувача з системою не викликає дискримінаційних результатів. Щоб запобігти упередженості на всіх рівнях, потрібно зробити кілька важливих кроків - від особистого використання моделі до її навчання та мінімізації упередженості:

- **Уникати явно дискримінаційних інструментів.** Слід уникати систем ШІ, розроблених компаніями, які відомі своєю дискримінаційною практикою або репутацією у суспільстві. Так само слід уникати систем ШІ, пов'язаних із скандалами та негативною реакцією громадськості через їх упередженість (*наприклад*, система онлайн-реклами Google показувала високооплачувані посади чоловікам частіше, ніж жінкам);
- **Фільтрувати упереджені навчальні дані.** Використовуючи публічно відкриті системи ШІ або системи, створені на замовлення ОГС, слід переконатися, що навчальні бази даних не відображають суспільні упередження та дискримінацію. Варто перевірити навчальні дані щодо стереотипів і провести тестування системи ШІ, щоб виявити упередженість у результатах. Будь-яка виявлена упередженість повинна бути виключена з бази даних;
- **Дотримуватись принципу репрезентації даних**. Для захисту від упередженості життєво необхідно переконатися, що дані є репрезентативними та різноманітними, інакше виникне упередженість через те, що люди певних груп не ідентифікуються. За потреби може знадобитися додаткова вибірка недостатньо представлених груп або генерування синтетичних даних, щоб уникнути упередженості щодо меншин, уразливих і маргіналізованих спільнот;
- **Налаштувати інструменти ШІ відповідно до місцевого контексту.** Часто системи ШІ, створені з певних джерел і з певними навчальними даними, можуть бути не в змозі налаштувати свої вхідні дані відповідно до місцевого контексту користувача, особливо якщо розробник не здійснює діяльність на цьому ринку. Щоб запобігти цьому, база даних та/або вхідні дані мають бути відібрані прикладами, які відповідають місцевому контексту;
- **Встановити фільтри запобігання упередженості для систем ШІ до їх вільного доступу до Інтернету**. Оскільки Інтернет містить необмежені джерела упереджених навчальних даних, життєво важливо запобігти навчанню систем ШІ за допомогою таких даних, забезпечивши їх фільтрами проти упередженості;
- **Забезпечити людський нагляд.** Щоб запобігти будь-яким упередженням у вихідних даних, самі користувачі повинні налаштувати результати моделей ШІ, щоб запобігти будь-яким культурним, расовим, гендерним чи будь-яким іншим упередженням у тексті. Регулярний моніторинг діяльності систем ШІ дозволяє ОГС вчасно виправляти виявлені в них упередження. Дізнайтеся більше про цю тему **в Розділі 6 цього Посібника (сторінка 23);**
- **Формувати запити без упередженості і дискримінаційних компонентів.** Через те, що моделі ШІ часто навчаються на вхідних даних користувачів, чим менше упередженості і дискримінації буде

відбуватися – тим кращі результати дає модель ШІ. А саме, чим більш конкретний опис надає користувач, тим менше стереотипів модель ШІ включає в створення зображення або тексту, і навпаки. Тому вкрай важливо уникати загальних описів та/або запитів, які самі по собі містять упередженість або стереотипне мислення;

- **Попереджати про потенційну упередженість у вихідних даних**. Публікуючи будь-який контент, на який впливали системи ШІ, потрібно зазначати про вплив ШІ на контент і попереджати аудиторію, що цей вплив може спричинити упередженість і неточності;
- **Оновлювати ШІ на основі скарг щодо упередженості та дискримінації.** Якщо ОГС отримує скаргу про те, що її система ШІ є упередженою або дискримінаційною, вона повинна розглянути таку скаргу, встановити причину інциденту та належно оновити систему ШІ, щоб вирішити проблему.

РОЗДІЛ 4. СВОБОДА ВИРАЖЕННЯ ПОГЛЯДІВ

Розповсюдження контенту з боку ОГС, створеного або модифікованого системами ШІ, може містити певні загрози щодо неправдивої або оманливої інформації, конфіденційної інформації та прозорості для аудиторії у її взаємодії з системами ШІ. Наприклад, дедалі популярніша діпфейк-технологія, цифрові аватари та чат-боти часто використовуються для поширення дезінформації у великих масштабах. Небезпечне використання цих та подібних технологій відкриває можливості для дезінформації, шахрайства, витоку конфіденційної інформації тощо.

Рекомендації. Щоб запобігти всім вищезазначеним ризикам, важливо забезпечити наявність фільтрів, маркування та перевірок ШІ-згенерованого контенту, і встановити додаткові механізми обробки такого контенту. Кілька кроків допомагають уникнути шкоди від ШІ-згенерованого контенту, пріоритезації та кураторства системами ШІ:

- **Розробляти політики конфіденційності.** ОГС має встановити та підтримувати політики конфіденційності, щоб забезпечити чіткі інструкції щодо поводження з конфіденційною інформацією. Дізнайтеся більше про це в **Розділі 2 цього Посібника (сторінка 10)**;
- **Маркувати контент, створений ШІ/модифікований ШІ.** Будь-який контент, створений або модифікований ШІ, потрібно позначати, щоб аудиторія ніколи не була введена в оману. Деякі системи ШІ або

вбудовані функції ШІ можуть автоматично надавати опції маркування, але доцільно завжди ініціювати маркування самостійно відповідно до вимог багатьох нормативних актів, національного законодавства та правил соціальних мереж;

- **Забезпечити людський нагляд.** Щоб запобігти будь-якій дезінформації у вихідних даних, користувачі самі повинні переглядати результати моделей ШІ, щоб редагувати фактично неправильну або оманливу інформацію з тексту. Дізнайтеся більше про це в **Розділі 6 цього Посібника (сторінка 23)**;
- **Перевіряти на достовірність.** Будь-який результат, створений системою ШІ, потрібно перевіряти на його відповідність оригінальному джерелу та меті його створення. Якщо система ШІ цитує будь-які джерела, їх також слід проаналізувати, перш ніж будь-яким чином використовувати вихідні дані системи;
- **Вимкнути автоматичне поширення/публікацію.** Автоматична публікація на кількох онлайн-платформах (включно з веб-сайтом ОГС) може призвести до розповсюдження небажаної, невідредагованої чи невідфільтрованої інформації, яку потім стає важче оцінити на кількох платформах одночасно. Отже, цю функцію потрібно деактивувати на всіх можливих платформах;
- **Використовувати надійні джерела з білого списку.** Слід створити список надійних новин і джерел, які можна використовувати як порівняльний показник того, чи є будь-яка інформація правдивою. Наприклад, можна використовувати кілька відомих газет, які заявляють про відповідність найкращим журналістським стандартам і практикам, як-от New York Times, The Guardian, BBC тощо. Цей білий список необхідно регулярно оцінювати та оновлювати у разі виникнення будь-яких розбіжностей, пов'язаних із білим списком медіа. Важливо також зазначити, що білі списки не можна сприймати як підставу вважати інформацію надійною за замовчуванням. Матеріали з таких джерел перевіряються, якщо виявлено найменші сумніви щодо вхідних даних, введених ШІ;
- **Проводити навчання з медіа грамотності для команди ОГС.** Організація повинна навчати свій персонал основним принципам медіа грамотності, пояснювати та навчати призначений персонал оцінювати достовірність інформації в Інтернеті, виявляти та боротися з дезінформацією, застосовувати OSINT та інші відповідні технології для проведення ретельного дослідження тощо;
- **Відповідально використовувати системи ШІ у журналістських розслідуваннях.** Будь-які випадки використання систем ШІ для цілей розслідування повинні бути розкриті аудиторії звітів про розслідування. Більше того, будь-які результати розслідувальних систем

ШІ завжди повинні ретельно контролюватися командою людського нагляду та повторно перевірятися із застосуванням додаткових джерел, які не базуються на ШІ. Важливо переконатися, що всі дані є загальнодоступними та збираються законно. Крім того, слідчі повинні використовувати технологію VPN для захисту свого розслідування;

- **Відповідально використовувати інструменти підбору контенту.** Використання інструментів ШІ, які забезпечують підбір контенту, має бути адаптовано до потреб аудиторії та враховувати права людини. Наприклад, ОГС має надати користувачам можливість змінювати налаштування на веб-сайтах, де публікується контент, щоб вони могли визначити пріоритетність контенту відповідно до своїх потреб або відмовитися від автоматизованого підбору контенту ШІ;
- **Відповідально використовувати інструменти модерації.** Використовуючи будь-які інструменти модерації контенту, ОГС має переконатися, що вони ефективні у підході до усього контенту (наприклад, мова ненависті, насильство в Інтернеті, сексуально-образливий контент тощо) без надмірної цензури коментарів користувачів і тих, хто може поділитися своїми поглядами в спеціально відведених розділах, або упередженого ставлення до певних поглядів, особливо політики та суспільно значущих подій. Крім того, ОГС має своєчасно оновлювати саму систему ШІ та її навчальні дані, щоб вона завжди могла належним чином охоплювати нові резонансні події;
- **Відповідально використовувати чат-ботів.** Під час використання чат-ботів ОГС повинна переконатися, що її співробітники не надають боту чутливу інформацію. Водночас навчання та розробка таких систем ШІ мають бути прозорими щодо використовуваних навчальних даних, і мають бути гарантовані фільтри проти шкідливої інформації чи упередженості, а також мають бути повідомлення для користувачів про те, що вони взаємодіють із ШІ, а не з реальною особою;
- **Відповідально використовувати цифрових аватарів.** Незважаючи на свої переваги, така технологія відкриває можливості для багатьох проблем, пов'язаних із крадіжкою особистості і шахрайством, тому щоразу, коли використовуються цифрові аватари, аудиторія має бути завжди повідомлена про те, що вони стикаються з аватаром, створеним ШІ, а не з реальною людиною. Під час навчання або розробки цих систем ШІ використання будь-якої персональної інформації як під час навчання, так і самим аватаром має завжди здійснюватися за обопільною згодою;
- **Своєчасно виявляти та відфільтровувати діпфейки.** Важливо застосовувати методи перевірки до підозрілого контенту, щоб перевірити діпфейки, розпізнаючи невідповідності голосу чи обличчя, аномалії кольору тощо. Нижче наведені приклади методів перевірки:

- **Аналіз рухів обличчя**. Виявляє підозрілі паттерни в анімації обличчя;
- **Аналіз текстури.** Виявляє підозрілі паттерни в текстурі шкіри і волосся;
- **Аудіо аналіз**. Виявляє розбіжності в синхронізації губ, якості звуку та частоті;
- **Аналіз метаданих**. Перевіряє метадані відео, наприклад, дату і час створення та географічне розташування.
- **Аналіз рівня помилок**. Аналізує кілька частот відео, щоб виявити неточності.

Крім того, будь-яка аудіальна інформація, яку обробляють ОГС, має бути перевірена через високий ризик підробки, оскільки аудіо-діпфейки найпростіше зробити. Доречно застосовувати кілька інструментів виявлення, щоб охопити найширший спектр і прояви діпфейків.

- **Використовувати автентифікацію.** Слід застосовувати методи автентифікації та відстежувати внутрішні індикатори у контенті, за допомогою яких можна виявити діпфейки. До таких маркерів належать:
  - **Цифрові водяні знаки:** Фрагмент цифрового коду чи зображення, вбудований у контент;
  - **Метадані:** Невід'ємні дані, що описують певний файл та/або фрагмент контенту;
  - **Блокчейн:** Відкрита технологія, яка використовує публічну прозорість як щит від діпфейків.

РОЗДІЛ 5. ЗАХИСТ ІНТЕЛЕКТУАЛЬНОЇ ВЛАСНОСТІ

Одним із життєво важливих питань, які виникають під час використання будь-якого програмного забезпечення на основі ШІ, є необхідність запобігти порушенню прав на об'єкти інтелектуальної власності (прав ІВ) як розробників програмного забезпечення на основі ШІ, так і інших користувачів. Особливо важливо забезпечити дотримання норм ІВ, якщо ОГС вирішить використовувати у своїй роботі унікальне програмне забезпечення. Проте не менш важливо забезпечити відповідальне використання загальнодоступних систем ШІ, оскільки інакше це може призвести до порушення прав ІВ та створити загрозу правових наслідків для організації.

Рекомендації. Щоб запобігти порушенню прав ІВ під час використання програмного забезпечення на основі ШІ, найефективнішим методом є застосування ліцензування програмного забезпечення до моделей, які

ОГС бажає використовувати, оскільки це визначає умови використання між ліцензіатом і розробником програмного забезпечення. Отже, ми рекомендуємо:

- **Уникати систем ШІ з порушенням прав ІВ**. Використання систем ШІ, які за замовчуванням порушують права ІВ, може призвести до відповідальності організації. Тому найкращим рішенням буде уникати використання таких систем, із відповідальним підходом до їх вибору. Дізнайтеся більше про це в **Розділі 7 цього Посібника (сторінка 30)**;
- **Уважно аналізувати застосовне законодавство.** Залежно від країни походження системи ШІ, умов ліцензійної угоди (якщо це актуально) і країни походження ОГС, до діяльності організації можуть застосовуватися різні національні закони. Основне законодавство, яке необхідно взяти до уваги, включає:
  - Міжнародні стандарти прав людини;
  - Міжнародні акти та стандарти з прав ІВ;
  - Національне законодавство про ІВ для країни діяльності ОГС;
  - Національне законодавство юрисдикції, зазначеної регулюючою в ліцензійній угоді, або в умовах використання (якщо це актуально);
  - Національне законодавство країни походження розробника системи ШІ (якщо це актуально);
- **Враховувати, що захищено ІВ**. При оцінці будь-яких вхідних чи вихідних даних системи ШІ необхідно брати до уваги типи робіт, які захищаються правилами ІВ. Наприклад, будь-які оригінальні ідеї, проекти, відкриття, винаходи та творчі роботи, створені окремою особою чи групою, захищені правами ІВ;
- **Дотримуватись доктрини добросовісного використання.** Ця доктрина дозволяє використовувати контент, захищений авторським правом, без дозволу власника прав за дотримання певних умов. Наприклад, якщо використання є некомерційним, проводиться для дослідження та призводить до нового іншого твору, такий акт, швидше за все, вважатиметься добросовісним використанням. Однак у кожному окремому випадку необхідно перевіряти, чи застосовуються правила добросовісного використання;
- **Використовувати ліцензовані моделі.** Замовляючи або налаштовуючи існуючу систему ШІ для використання, ОГС повинна укласти ліцензійну угоду з розробником, щоб запобігти порушенням прав ІВ для обох сторін. Те саме стосується будь-якого зовнішнього програмного забезпечення на основі ШІ, яке використовує ОГС і яке не було розроблено самою ОГС;
- **Дотримуватись ліцензійних угод.** Укладаючи ліцензійну угоду, організація повинна ретельно проаналізувати положення такої угоди,

особливо те, як вони визначають, хто володіє правами на систему ШІ, результати її роботи, програмне забезпечення, дані тощо, і хто несе відповідальність за порушення прав ІВ згідно з угодою. Важливо дотримуватися цих положень після укладання угоди;

- **Обирати вхідні дані в моделях ШІ відповідно до стандартів ІВ**. Користувач повинен належним чином розробляти запити, щоб переконатися, що введення в модель не призведе до результатів, які порушують чиїсь права. Ось два способи мінімізувати таку ймовірність:

Загальний підхід:	Введення у модель ШІ якомога меншої кількості загальної інформації.
Особливий підхід:	Створення запиту оригінальним та настільки конкретним і детальним, наскільки можливо.

- **Перевіряти відповідність вихідних даних стандартам ІВ.** Вихідну інформацію моделі ШІ можна захистити від порушення прав ІВ двома основними кроками:

Перевірка посилання:	Здійснення пошуку оригінального контенту в інтернеті, щоб переконатися у відсутності порушення авторського права
Модифікація вихідних даних:	Змінення результатів, отриманих від моделі ШІ, щоб зробити кінцевий продукт унікальним контентом і запобігти порушенням..

- **Захистити контент, створений ШІ ОГС, від претензій третіх сторін/ порушень прав ІВ.** Найкращий спосіб захистити контент ОГС від будь-яких судових позовів — це відповідально використовувати системи ШІ та гарантувати, що під час створення такого контенту не буде порушено ІВ, як зазначено вище. Щоб захистити контент ОГС від порушень прав ІВ, організація повинна публічно вказати, що вона є правовласником цього контенту, і контент має бути захищений патентом або комерційною таємницею, якщо це актуально. Така ж тактика застосовується до захисту унікальних функцій системи ШІ, повністю розробленої ОГС.

РОЗДІЛ 6. ЛЮДСЬКИЙ НАГЛЯД

Щоб гарантувати, що ОГС успішно дотримується всіх відповідних правил і стандартів, а також безпечно та продуктивно використовує моделі ШІ, важливо забезпечити людський нагляд. Будь-яка діяльність, яка пов'язана з використанням моделей ШІ, повинна включати механізми людського нагляду, щоб запобігти будь-яким шкідливим наслідкам, які можуть спричинити моделі ШІ.

Рекомендації. Щоб запобігти будь-яким неточностям вихідних даних моделей ШІ, необхідний людський нагляд за діяльністю моделей ШІ. Не тільки користувачі повинні перевіряти результати перед їх використанням, але й сама організація повинна запровадити політики, навчити команду та постійно здійснювати моніторинг діяльності систем ШІ, особливо якщо ОГС навчає свою власну модель ШІ або використовує ліцензоване програмне забезпечення. Ось кілька важливих кроків, які необхідно зробити, щоб підтримувати необхідний рівень людського нагляду:

- **Призначити відповідальну особу (осіб).** Для організованого здійснення людського нагляду життєво важливо призначити ключових членів команди, відповідальних за людський нагляд, і розподілити між ними функції моніторингу, оцінки та прийняття рішень;
- **Навчати команду ОГС.** Усі члени команди, які мають взаємодіяти з системами ШІ, повинні пройти навчання з питань політики організації щодо систем ШІ та загальних правил відповідального використання ШІ. Вони повинні знати про всі протоколи дій у кризових ситуаціях та знати, як безпечно використовувати системи ШІ, особливо ті, що розроблені, замовлені або налаштовані ОГС. Окремі відділи ОГС можуть бути навчені використовувати різні системи залежно від їхніх потреб і сфери відповідальності;
- **Забезпечувати зворотний зв'язок.** Слід запровадити регулярні цикли зворотного зв'язку всередині команди щодо роботи моделі ШІ, що допоможе скоригувати вхідні дані та політики щодо ШІ, яку підтримує організація;
- **Розробити протоколи дій у кризових ситуаціях.** ОГС повинні створювати та періодично оновлювати протоколи дій у кризових ситуаціях, які чітко визначають обов'язки команди під час реагування на кризу, а також надають інструкції щодо ефективної комунікації у кризових ситуаціях і конкретні тактики щодо безпекової, технічної, репутаційної та інших видів шкоди. Такі протоколи повинні бути завжди доступними та добре відомими команді;

- **Розробити портал для скарг.** ОГС, яка публікує будь-який контент, повинна створити механізм подання скарг, який дозволить аудиторії повідомляти про проблеми та забезпечувати зворотний зв'язок з ОГС. Організація повинна призначити персонал для моніторингу та відповідей аудиторії через портал скарг;
- **Регулярно оновлювати систему.** ОГС, яка розробляє, налаштовує або замовляє системи ШІ, повинна проводити регулярні оновлення, щоб забезпечити кращий захист і простіший, більш скоординований робочий процес команди людського нагляду. Так само використання загальнодоступних систем ШІ має включати регулярний контакт із розробником для оновлень системи ШІ;
- **Регулярно проводити аналіз дотримання прав людини.** Людський нагляд передбачає регулярне проведення ОВПЛ та аналіз рівня дотримання організацією прав людини. Відповідальні особи повинні надавати регулярні звіти щодо ОВПЛ та успішного людського нагляду. Дізнайтеся більше про це в **Розділі 1 цього Посібника (сторінка 6)**.

РОЗДІЛ 7. КЕРІВНИЦТВО ДЛЯ ВІДПОВІДАЛЬНОГО ВИБОРУ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ

Щоб правильно обрати інструменти ШІ, ОГС має зібрати інформацію про них і перевірити, чи відповідають вони критеріям безпечних інструментів ШІ з низьким рівнем ризику. Для зручності процесу обрання, Лабораторія цифрової безпеки розробила керівництво для кожного розділу цього Посібника, яке ОГС можуть використовувати, щоб переконатися, що інструмент ШІ справді безпечний у використанні. Алгоритми використання керівництва наступні:

1. У кожному розділі дайте відповідь «**ні**» або «**так**» на питання в лівій колонці;
2. Якщо ваша відповідь відповідає умовам рекомендації (права колонка), дійте відповідно до цієї рекомендації.

Використовуючи це керівництво, слід взяти до уваги три важливі зауваження:

- Рядки для запитань і рекомендацій у кожному розділі забарвлені **червоним**, **жовтим** або **зеленим** кольором залежно від рівня ризику. Червоний рядок позначає неприйнятний ризик, жовтий рядок позначає середній ризик, а зелений ряд позначає низький ризик або його відсутність (важливі моменти, які слід враховувати);
- Щоб правильно оцінити безпеку певних інструментів ШІ, важливо перевірити та відповісти на запитання в усіх розділах, наведених нижче;
- Важливо зазначити, що це керівництво не є вичерпним, і можуть існувати інші питання, які вимагають уваги ОГС, залежно від характеру інструменту ШІ та специфіки роботи ОГС. Організація повинна завжди залишатися пильною щодо нових ризиків і разом із цим Посібником використовувати додаткові інструменти перевірки відповідності.

Оцінка впливу на права людини та оцінка ризиків

Питання	Ні	Так	Рекомендація
Чи розроблена система ШІ компанією зі штаб-квартирою в країні з низьким індексом захисту прав людини (наприклад, Росія, Китай, Іран)?	☐	☐	Якщо «**так**», цей інструмент ШІ становить неприйнятний ризик і не рекомендований до використання.
Чи гарантує розробник або постачальник систем ШІ дотримання всіх відповідних правових норм?	☐	☐	Якщо «**ні**», цей інструмент ШІ створює неприйнятний ризик і не рекомендується для використання, доки не будуть надані такі гарантії (розкриті звіти про відповідність тощо).
Чи розроблено систему ШІ з порушенням прав людини (наприклад, скрейпінг даних або явно дискримінаційні алгоритми)?	☐	☐	Якщо «**так**», цей інструмент ШІ становить неприйнятний ризик і не рекомендований до використання..
Чи має розробник чи постачальник системи ШІ сумнівну репутацію або відомий масовими порушеннями прав людини?	☐	☐	Якщо «**так**», це вимагає додаткової уваги з боку команди ОВПЛ/ОР та вжиття заходів для забезпечення використання конкретної системи ШІ з дотриманням прав людини.
Чи містять умови використання сумнівні положення щодо платежів, прав ІВ на створений контент, відповідальності, безпеки, конфіденційності тощо?	☐	☐	Якщо «**так**», це вимагає додаткової уваги з боку команди ОВПЛ/ОР та спілкування з розробником для уточнення цих положень. Якщо положення неможливо уточнити, доцільно утриматися від використання цієї системи ШІ.
Чи передбачається умовами використання, що користувач несе відповідальність за порушення прав ІВ, порушення безпеки даних, поширення дезінформації тощо?	☐	☐	Якщо «**так**», використання цього інструменту ШІ створює дуже великий тягар на ОГС, і команда ОР повинна оцінити, чи можливо використовувати такий інструмент ШІ за цих обставин.
Чи надає розробник механізм сповіщень або портал скарг для свого інструменту ШІ?	☐	☐	Якщо «**ні**», ОГС має створити механізм сповіщення розробника про проблеми чи скарги щодо системи ШІ.
Чи проводить розробник регулярні ОВПЛ/ОР та/чи інші процеси, щоб уникнути або пом'якшити несприятливий вплив своєї системи ШІ?	☐	☐	Якщо «**ні**», ОГС повинні проявити належну обачність і провести більш ретельну ОВПЛ/ОР. Якщо це неможливо, доцільно утриматися від використання цієї системи ШІ.

Цифрова безпека та захист персональних даних

Питання	Ні	Так	Рекомендація
Чи відомий цей інструмент ШІ скрейпінгом даних, незаконним збором і обробкою персональних і конфіденційних даних або іншою незаконною діяльністю?	☐	☐	Якщо «**так**», цей інструмент ШІ становить неприйнятний ризик і не рекомендований до використання.
Чи має розробник системи ШІ історію добросовісного вирішення інцидентів безпеки та постійного вдосконалення своїх заходів безпеки проти витоку даних?	☐	☐	Якщо «**ні**», цей інструмент ШІ становить неприйнятний ризик і не рекомендований до використання.
Чи дозволяє система ШІ своєчасно видаляти конфіденційну інформацію?	☐	☐	Якщо «**ні**», ОГС має переконатися, що жодна конфіденційна інформація ніколи не передається інструменту ШІ (наприклад, через запити).
Чи мають суб'єкти даних право заперечувати проти використання їхніх ПД інструментом ШІ та, якщо необхідно, отримати до них доступ і видалити їх?	☐	☐	Якщо «**ні**», ОГС має переконатися, що використання системи ШІ завжди базується на законних підставах і є прозорим для суб'єктів даних.
Чи цей інструмент ШІ натренований на конфіденційних даних?	☐	☐	Якщо «**так**», розробник/постачальник повинен гарантувати ОГС, що ці дані були зібрані законним шляхом і не будуть розголошені під час використання інструменту ШІ. В іншому випадку цей інструмент ШІ створює неприйнятний ризик і не рекомендується для використання.
Чи дозволяє інструмент ШІ персоналізувати налаштування, наприклад, заборонити використання ПД для навчання, відмовитися від зберігання даних у хмарі чи заборонити доступ до файлів на пристрої?	☐	☐	Якщо «**ні**», це може становити значний ризик, якщо розробник не надасть достатніх гарантій безпеки від витоку даних. Радимо ОГС уникати надання конфіденційної інформації таким системам ШІ.
Чи може інструмент ШІ будь-коли збирати аудіальну інформацію шляхом фонового прослуховування, просто встановивши його на пристрій?	☐	☐	Якщо «**так**», то ОГС має врахувати це та вжити додаткових заходів безпеки під час використання такого інструменту.
Чи має розробник/постачальник інструменту ШІ політику захисту даних?	☐	☐	Якщо «**ні**», ОГС повинна взяти це до уваги і вирішити усі відповідні питання з розробником/провайдером.

Запобігання алгоритмічній упередженості і дискримінації

Питання	Ні	Так	Рекомендація
Чи відомий цей інструмент ШІ явно дискримінаційними практиками чи асоціюється із суспільним невдоволенням чи скандалами через свою упередженість?	☐	☐	Якщо «**так**», цей інструмент ШІ становить неприйнятний ризик і не рекомендований до використання.
Чи був цей інструмент ШІ навчений із використанням неупереджених, різноманітних і репрезентативних даних?	☐	☐	Якщо «**ні**», цей інструмент ШІ становить неприйнятний ризик і не рекомендований до використання.
Чи гарантує розробник запобіжні заходи проти упередженості та/чи алгоритмічних галюцинацій?	☐	☐	Якщо «**ні**», ОГС має вирішити цю проблему за допомогою покращеної модерації результатів інструменту ШІ.
Чи можна налаштувати цей інструмент ШІ відповідно до місцевого контексту кожної країни чи культури, що його використовує?	☐	☐	Якщо «**ні**», цей інструмент ШІ може спричинити проблеми для ОГС, які працюють у контекстах, що не входять до сфери навчальних даних інструменту ШІ. Краще додатково перевірити результати цього інструменту або уникати його використання.
Чи встановлено в інструменті ШІ фільтри запобігання упередженості, коли він має доступ до Інтернету?	☐	☐	Якщо «**ні**», існує висока ймовірність того, що навчальні дані зіпсовані, і використання такого інструменту слід або уникати, або належним чином контролювати. Або ОГС має встановити такі фільтри самостійно.
Чи вчасно оновлюються навчальні дані інструменту ШІ, наприклад, після скарг або інцидентів?	☐	☐	Якщо «**ні**», нові інциденти або проблеми, які не охоплені інструментом ШІ, підвищать ризик неправильних результатів, тому краще уникати таких інструментів ШІ.
Чи можуть користувачі або ОГС отримати доступ до навчальних даних і налаштувати їх?	☐	☐	Якщо «**ні**», розробник повинен гарантувати відсутність упередженості в інструменті ШІ та своєчасне оновлення навчальних даних.

Свобода вираження поглядів

Питання	Ні	Так	Рекомендація
Чи забезпечує інструмент ШІ автоматичний обмін/публікацію згенерованих результатів?	☐	☐	Якщо «**так**», має бути можливість відмовитися від цієї функції, інакше цей інструмент ШІ становить неприйнятний ризик і не рекомендується для використання.
Чи встановлено в інструменті ШІ фільтри проти дезінформації та упередженої чи ненависницької інформації?	☐	☐	Якщо «**ні**», цей інструмент ШІ становить неприйнятний ризик і не рекомендований до використання.
Чи надає інструмент ШІ можливість позначати ШІ-згенерований контент як такий?	☐	☐	Якщо «**ні**», цей інструмент ШІ створює неприйнятний ризик, якщо тільки ОГС сама не позначає контент як створений ШІ або модифікований ШІ, залежно від випадку.
Чи встановлює інструмент ШІ автоматичний обмін/публікацію разом зі своєю роботою?	☐	☐	Якщо «**так**», має бути можливість відмовитися, інакше цей інструмент ШІ становить неприйнятний ризик і не рекомендується для використання.
Чи регулярно інструмент ШІ оновлює дані?	☐	☐	Якщо «**ні**», ОГС має вжити додаткових заходів, щоб переконатися, що інструмент ШІ актуалізований. Наприклад, звернутись до розробника/постачальника або змінити інструмент ШІ (якщо можливо).
Чи дозволяє інструмент кураторства контенту користувачам налаштовувати свої параметри відповідно до]їхніх уподобань?	☐	☐	Якщо «**ні**», такі інструменти кураторства контенту слід або оновити, щоб забезпечити належне налаштування, або ОГС слід уникати їх використання.
Чи забезпечує інструмент модерації відповідний фільтр без упередженості чи непотрібної критики?	☐	☐	Якщо «**ні**», навчальні дані або сам інструмент ШІ потрібно оновити, щоб він був ефективним, інакше ОГС має уникати цього.
Чи чат-бот одночасно застосовує достатні фільтри проти шкідливої інформації та не навчений на конфіденційних даних?	☐	☐	Якщо «**ні**», інформаційні фільтри потрібно виправити або оновити, а розробник повинен гарантувати, що використання конфіденційних навчальних даних є законним. В іншому випадку такий чат-бот не рекомендується використовувати.
Чи цифровий аватар навчався з ПД персоналу ОГС?	☐	☐	Якщо «**так**», мають бути гарантії того, що збір таких даних завжди відбувається за згодою та недоступний для третіх сторін, включаючи розробників/ постачальників системи ШІ.
Чи надає інструмент ШІ цитати та джерела зі створеним контентом?	☐	☐	Якщо «**ні**», повна перевірка фактів щодо результатів інструменту ШІ має бути виконана людиною-наглядачем.

Права ІВ

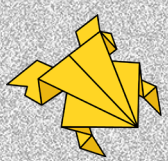
Питання	Ні	Так	Рекомендація
Чи відомий цей інструмент ШІ тим, що порушує авторські права чи інші права ІВ інших людей?	☐	☐	Якщо «**так**», цей інструмент ШІ становить неприйнятний ризик і не рекомендований до використання..
Чи є ліцензійна угода чіткою, прозорою та чи визначає вона всі права та обов'язки обох сторін?	☐	☐	Якщо «**ні**», слід уникати укладення такої ліцензійної угоди, натомість ОГС може запропонувати зміни до угоди для вирішення та коригування проблем.
Чи гарантує розробник відсутність порушень прав ІВ як під час навчання, так і під час використання цього інструменту ШІ?	☐	☐	Якщо «**ні**», для ОГС існує високий ризик нести спільну відповідальність за порушення прав ІВ, тому такий інструмент ШІ не рекомендується використовувати.
Чи передбачає розробник, що права на контент, створений або змінений за допомогою цього інструменту ШІ, належать кожному окремому користувачеві?	☐	☐	Якщо «**ні**», ОГС має забезпечити додаткову модифікацію створеного ШІ контенту, щоб перетворити його на оригінальний контент, захищений авторським правом.
Чи вчасно оновлюється інструмент ШІ після запровадження нових правил щодо ІВ?	☐	☐	Якщо «**ні**», ОГС має проаналізувати відповідність такого інструменту ШІ оновленим правилам щодо ІВ. Якщо необхідний рівень відповідності не досягнуто, краще уникати використання цього інструменту ШІ.
Чи можна використовувати створений ШІ контент у комерційних, творчих чи законних цілях?	☐	☐	Якщо «**ні**», ОГС має врахувати це як один із факторів у процесі відбору залежно від мети використання системи ШІ.

ВИСНОВКИ

Після розгляду всіх наведених рекомендацій життєво важливо зважити та збалансувати інтереси ОГС при використанні будь-яких інструментів ШІ. Зокрема, у більшості випадків остаточна відповідальність за несприятливі наслідки, спричинені безвідповідальним використанням систем ШІ, лежить на самій ОГС, особливо, якщо ОГС розробила, замовила або налаштувала систему ШІ, а не використовує загальнодоступний інструмент. Тому вкрай важливо застосовувати відповідні стандарти для ШІ на всіх етапах життєвого циклу системи ШІ – від ранньої розробки до використання та припинення системи.
Перш ніж розпочати впровадження будь-яких інструментів ШІ у свою роботу, ОГС має ретельно оцінити ризики та переваги систем і вирішити, чи потенційні переваги переважають ризики та небезпеки. У всіх випадках застосування систем ШІ, ОГС повинні дотримуватися **основних принципів відповідального використання ШІ**, таких як:

- прозорість і підзвітність,
- безпека даних і захист конфіденційної інформації,
- ефективний та професійний людський нагляд,
- орієнтоване на права людини використання інструментів на основі ШІ.

Крім того, ОГС має завжди **слідкувати за останніми оновленнями в регуляторній сфері**, належним чином впроваджувати нові законодавчі ініціативи та стандарти, а також перевіряти відповідність своєї практики міжнародним стандартам. Наостанок, однією з ключових особливостей у сфері ШІ є відповідальний вибір систем ШІ, особливо тих, які призначені для корпоративного, а не особистого використання. Впроваджуючи інструменти ШІ на організаційному рівні, ОГС, як суб'єкти вразливої та ризикованої природи, повинні ретельно обирати розробників і виробників систем ШІ та повторно перевіряти їх вплив за допомогою добре розроблених процедур ОВПЛ та ОР. Прогрес неможливої не слід зупиняти, але ми можемо змусити його працювати на благо громадянського суспільства, діючи сумлінно та відповідально!



Лабораторія
Цифрової
Безпеки