ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД
«УНІВЕРСИТЕТ ЕКОНОМІКИ ТА ПРАВА «КРОК»
Кафедра журналістики

Гергелюк Анастасія Андріївна

БАКАЛАВРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

« Роль штучного інтелекту в боротьбі з дезінформацією »

(тема)

061 Журналістика
(шифр і назва спеціальності)

«Журналістика»
(освітня програма)

Подається на здобуття освітнього ступеня бакалавр

Бакалаврська кваліфікаційна робота **Гергелюк Анастасії Андріївни** містить результати власних доробок. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____
(підпис, ініціали та прізвище здобувача)

Науковий керівник: **Хоменко Ангеліна Олександрівна**
Викладач кафедри журналістики

Київ - 2025р.

ЗМІСТ

ВСТУП

Актуальність теми

В умовах цифрової епохи дезінформація стає серйозною загрозою для суспільства, демократії та безпеки держав. Завдяки розвитку соціальних мереж і цифрових платформ поширення фейкових новин, маніпулятивного контенту та інформаційних атак відбувається з безпрецедентною швидкістю. Традиційні методи виявлення дезінформації виявляються недостатньо ефективними, тому зростає потреба у використанні сучасних технологій, зокрема штучного інтелекту.

ШІ пропонує нові можливості для аналізу інформації, виявлення фальшивих новин та нейтралізації деструктивного впливу маніпулятивного контенту. Використання алгоритмів машинного навчання, нейромереж та обробки природної мови дозволяє ідентифікувати неправдиві повідомлення, аналізувати їхній вплив і навіть прогнозувати майбутні загрози. Проте застосування ШІ у цій сфері стикається з низкою викликів, включаючи етичні питання, можливість алгоритмічних помилок та ризик використання цих технологій для цензури.

Дослідження ролі штучного інтелекту у боротьбі з дезінформацією є актуальним не лише для світової наукової спільноти, а й для українського інформаційного простору. В умовах гібридної війни Україна стала одним із основних об’єктів інформаційних атак, що підкреслює необхідність розробки ефективних методів протидії. Дослідження у цій сфері сприятиме вдосконаленню інформаційної політики держави, розвитку медіаграмотності та створенню ефективних алгоритмів протидії фейкам.

Об’єкт дослідження – механізми боротьби з дезінформацією в сучасному інформаційному просторі. Це охоплює всі аспекти протидії маніпуляціям, фальшивим новинам та ворожій пропаганді, що активно поширюються через медіа та соціальні мережі в умовах цифрової епохи.

Предмет дослідження – це використання технологій штучного інтелекту в журналістиці для виявлення, верифікації та протидії дезінформації в сучасному медіапросторі.

Ступінь наукової розробки теми

Тема боротьби з дезінформацією за допомогою штучного інтелекту є предметом досліджень як українських, так і закордонних учених.

Серед зарубіжних дослідників значний внесок у вивчення механізмів протидії дезінформації зробили:

H. Allcott і M. Gentzkow у своїй роботі "Social Media and Fake News in the 2016 Election" дослідили вплив соціальних мереж на поширення фейкових новин під час виборів у США, а також оцінили ефективність методів їх виявлення.

C. Silverman досліджував алгоритмічні маніпуляції контентом у медіа та технології перевірки фактів, що використовуються для боротьби з дезінформацією.

H. Shu, S. Sliva, J. Wang та інші у дослідженні "Fake News Detection on Social Media: A Data Mining Perspective" розглянули підходи машинного навчання для автоматичного розпізнавання фейкових новин.

В Україні питання інформаційної безпеки та методів боротьби з дезінформацією досліджуються як науковцями так і державними установами. Ось деякі з них:

Науковці:

У. Ільницька – досліджує проблему інформаційної безпеки України та захисту національного інформаційного простору від негативних пропагандистсько-маніпулятивних впливів. Наукові Журнали та Конференції

О. Онищенко, О. Панченко, Я. Гмир – аналізують основні загрози інформаційній безпеці та механізми протидії дезінформації. Наукова періодика КУ

Б. Миколайчук – виокремлює критерії оцінки дезінформації та основні шляхи її протидії. Наукова періодика КУ

В. Лізунова – аналізує гібридні інформаційні війни в умовах цифрової епохи.

Державні установи:

Центр стратегічних комунікацій та інформаційної безпеки – це урядова організація, створена при Міністерстві культури та інформаційної політики України, яка займається аналізом інформаційного середовища, моніторингом дезінформаційних наративів та підвищенням обізнаності громадян про інформаційні загрози. https://spravdi.gov.ua/

Попри наявність наукових праць, питання ефективності штучного інтелекту в боротьбі з дезінформацією потребує подальших досліджень, особливо у контексті його застосування в українських реаліях. Недостатньо вивченими залишаються аспекти точності алгоритмів у визначенні маніпулятивного контенту, та проблема етичності їхнього застосування, та шляхи удосконалення ШІ у боротьбі з інформаційними загрозами.

Мета і завдання дослідження

Метою дослідження є – аналіз можливостей та викликів використання штучного інтелекту для виявлення та протидії дезінформації, а також розробка рекомендацій щодо його ефективного впровадження в інформаційний простір України.

Для досягнення цієї мети необхідно вирішити такі завдання:

Теоретико-методологічні основи використання штучного інтелекту в медіа:

- Дослідити сутність та особливості дезінформації в сучасному цифровому середовищі, визначити її основні типи та механізми поширення.
- Охарактеризувати основні методи та технології штучного інтелекту, що застосовуються для аналізу інформації, їхню ефективність та обмеження.
- Провести огляд наукових підходів та сучасних досліджень у сфері боротьби з дезінформацією.

Аналіз використання штучного інтелекту в зарубіжних та вітчизняних медіа для протидії дезінформації

Проаналізувати сучасні алгоритми штучного інтелекту, які використовуються для боротьби з фейковими новинами, виявлення маніпулятивного контенту та перевірки фактів.

Дослідити практики використання ШІ в зарубіжних та вітчизняних медіа.

Провести SWOT-аналіз ефективності технологій ШІ у боротьбі з дезінформацією.

Рекомендації та пропозиції щодо вдосконалення використання штучного інтелекту у протидії дезінформації

Потрібно оцінити ефективність технологій ШІ у виявленні дезінформації, використавши аналітичні методи та реальні кейси застосування в медіа.

Встановити основні виклики та ризики використання ШІ у боротьбі з дезінформацією, включаючи проблеми етичності, точності алгоритмів та потенційну загрозу цензури.

Розробити рекомендації щодо вдосконалення застосування ШІ у сфері інформаційної безпеки України, враховуючи міжнародний досвід та сучасні тенденції розвитку цифрових технологій.

Методи дослідження

Для досягнення поставлених завдань у роботі застосовувалися такі методи дослідження:

Аналіз і синтез – для вивчення наукових джерел з питань дезінформації та штучного інтелекту.

Контент-аналіз – для дослідження інформаційного контенту, який класифікується як дезінформація.

Системний аналіз – для оцінки ефективності існуючих алгоритмів ШІ у боротьбі з фейковими новинами.

SWOT-аналіз – для визначення сильних і слабких сторін використання ШІ в інформаційній безпеці.

Порівняльний аналіз – для зіставлення методів протидії дезінформації в різних країнах.

Наукова новизна одержаних результатів

Новизна роботи полягає в комплексному аналізі можливостей та викликів застосування штучного інтелекту у боротьбі з дезінформацією, а також у розробці рекомендацій щодо його ефективного впровадження в інформаційний простір України. У роботі вперше здійснено ґрунтовний SWOT-аналіз технологій ШІ у контексті інформаційної безпеки.

Практична значущість дослідження

Результати дослідження можуть бути використані для розробки державних стратегій інформаційної безпеки, створення ефективних алгоритмів виявлення дезінформації та вдосконалення медіа грамотності. Запропоновані рекомендації можуть бути корисними для журналістів, державних установ та розробників алгоритмів штучного інтелекту.

Структура та обсяг роботи

Дипломна робота складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатків. Загальний обсяг роботи становить 59 сторінок.

РОЗДІЛ 1. ДОСЛІДЖЕННЯ ТЕРМІНІВ І ПОНЯТЬ ДОСЛІДЖЕННЯ, ОГЛЯД ЛІТЕРАТУРИ

1.1. **Поняття дезінформації та її вплив на сучасне суспільство**

Дезінформація - це навмисне створення, модифікація або поширення неправдивої чи маніпулятивної інформації з використанням технологій штучного інтелекту, що має на меті ввести в оману користувача, викривити уявлення про події або вплинути на політичні, економічні, соціальні процеси.[1]

Дезінформація є однією з найбільших загроз для сучасного інформаційного середовища, зокрема в умовах цифровізації та швидкого розвитку нових медіа. Термін «дезінформація» вживається для позначення неправдивої, або спотвореної інформації, яка поширюється з метою введення в оману або маніпулювання громадською думкою[2]. Відмінність дезінформації від інших форм неправдивої інформації, таких як помилки чи непорозуміння, полягає в навмисному характері її створення та поширення в маси. На відміну від випадкових неточностей, або міфів, дезінформація має чітко визначену мету — вплив на когось або на аудиторію для досягнення політичних, економічних, соціальних або особистих цілей.

У сучасному світі дезінформація активно використовує можливості цифрових технологій, зокрема соціальних мереж та інтернет-платформ, що значно ускладнює її виявлення, та боротьбу з дезінформацією. Інтернет і соціальні мережі дають можливість поширювати неправдиву інформацію на неймовірно велику аудиторію за короткий проміжок часу. Однією з характерних рис дезінформації є її здатність миттєво охоплювати великі соціальні групи, завдяки чому вона стає потужним інструментом для впливу на суспільство.

Дезінформація може набувати різних форм, таких як фальшиві новини, меми, маніпуляції через соціальні мережі, створення фальшивих профілів або підроблених даних. [3]

У багатьох випадках, ці форми дезінформації мають схожі ознаки з традиційними, правдивими медіа-форматами, що ускладнює для користувачів розрізнення правдивої інформації від фальшивої. Сьогодні можна спостерігати активне поширення фальшивих новин, які часто маскують під справжні новини, що використовуються для зформовування хибних уявлень, або маніпулювання громадською думкою.

Поширення дезінформації не є безневинним процесом, вона має глибокий вплив на соціальні, політичні та економічні процеси, створюючи невизначеність і нестабільність. Однією з основних загроз, які виникають внаслідок дезінформації, є її здатність викликати соціальні конфлікти та напруження в суспільстві. Зокрема, на політичному рівні, дезінформація може використовуватися для зміни громадської думки про певних політичних лідерів чи партій, для підриву довіри до державних інститутів або для маніпулювання виборчими процесами. Історія знає чимало випадків, коли країни використовували дезінформацію для впливу на вибори в інших державах, що стало основою для нових соціальних і політичних потрясінь.

Крім того, дезінформація має безпосередній вплив на економіку. Фальшиві новини та маніпуляції можуть призводити до фінансових криз або змін у поведінці споживачів. Наприклад, поширення неправдивої інформації про компанії або продукти може призвести до значних фінансових втрат, коли люди приймають рішення, засновані на хибних фактах. У сфері охорони здоров'я дезінформація є особливо небезпечною, адже помилкові відомості про ліки, лікувальні методи чи вакцини можуть серйозно вплинути на здоров'я громадян, сприяти паніці або відмові від важливих медичних процедур.

Одним із найбільших викликів для сучасних демократичних суспільств є дезінформація, яка підриває довіру до традиційних медіа та інститутів влади. В

умовах глобалізації та розвитку цифрових технологій особливо актуальною стає проблема впливу на міжнародні відносини.

Дезінформація може бути інструментом гібридної війни, де інформаційні атаки використовуються для розколу суспільства в країні-цілі, для дестабілізації ситуації в певному регіоні або для дезорганізації діяльності міжнародних організацій.

Таким чином, дезінформація є серйозною загрозою для стабільності сучасного інформаційного простору. [4]

Її наслідки можуть бути руйнівними як для окремих осіб, так і для цілого суспільства, оскільки вона ставить під загрозу не лише правильність сприйняття подій, а й саму основу довіри в соціумі людей. У зв'язку з цим, боротьба з дезінформацією стає надзвичайно важливим завданням для урядів, медіа-організацій, а також для технологічних компаній і розробників інфориаційних рішень.

Зростання впливу дезінформації в цифрову епоху також змушує переглянути поняття інформаційної безпеки, інформаційна безпека в цьому контексті вже не обмежується лише захистом технічних систем, а включає захист інформаційного середовища від маніпуляцій, фейків та атак на когнітивну сферу особистості. Цей аспект часто описується терміном **«когнітивна безпека»**, що означає здатність суспільства протистояти впливу шкідливої інформації, зберігаючи здатність до критичного мислення та об’єктивного аналізу.

У сучасній науковій літературі також виділяють поняття **«інформаційної гігієни»** — тобто навички та звички, які допомагають користувачам орієнтуватися в інформаційному просторі, відокремлюючи достовірні джерела від дезінформації. Сюди входить перевірка фактів, вміння розпізнавати джерела інформації, розуміти природу алгоритмів соціальних мереж, які формують неправдивий інформаційний простір навколо користувача.

Іншою важливою характеристикою дезінформації є її **емоційний заряд**. На відміну від нейтральної або раціональної інформації, дезінформація часто спонукає до страхів, обурення чи ненависті. Це дозволяє їй швидше поширюватися через механізм вірусного контенту.

Як зазначають дослідники, такі емоційно забарвлені повідомлення на 70% частіше поширюються користувачами, ніж фактичні новини.

Варто також звернути увагу на феномен **«інформаційного шуму»**, коли велика кількість неправдивих або частково правдивих повідомлень створює відчуття хаосу, ускладнюючи пошук істини. Це стратегія, яку часто використовують в інформаційних війнах для деморалізації противника, розмивання кордонів між правдою та брехнею.

З теоретичної точки зору, дезінформацію можна розглядати крізь призму таких концепцій, як:

- **теорія масової комунікації**, що аналізує, як повідомлення передаються та сприймаються у суспільстві;
- **медіа-ефекти**, зокрема ефект фреймінгу, коли дезінформація змінює спосіб інтерпретації фактів через певну подачу;
- **теорія спіраль мовчання**, за якою домінування дезінформаційного дискурсу змушує менш популярні (але правдиві) позиції залишатися поза увагою.

У міжнародному контексті дезінформація все частіше стає елементом так званих **«гібридних загроз»**, що включають поєднання кібероперацій, економічного тиску, військових провокацій та інформаційних атак. Випадки втручання у вибори в США, Великій Британії (Brexit), Німеччині, Франції та багатьох інших країнах засвідчують, що дезінформація використовується як інструмент зовнішньополітичного впливу.

Додатково слід згадати про **роль автоматизованих бот-мереж** (ботоферм), які масово поширюють певні наративи, створюючи ілюзію громадської думки. Такі

інструменти активно застосовувалися у конфліктних ситуаціях, зокрема у війні Росії проти України, для просування пропагандистських меседжів, дискредитації Збройних сил України, міжнародних партнерів, волонтерів, а також для поширення паніки серед цивільного населення.

Із загостренням глобальної боротьби за інформаційний простір постає завдання **формування інформаційного імунітету** на рівні окремих громадян, освітніх установ та цілих держав. Це вимагає оновлення освітніх програм, запровадження медіаграмотності, підтримки незалежних медіа та розробки законодавчих механізмів для стримування дезінформаційних атак.

1.2. Поняття та розвиток штучного інтелекту

Процес використання роботів для імітації людського інтелекту з метою виконання різноманітних завдань, які раніше були властиві тільки людям називається - штучний інтелект [5]. У найширшому розумінні, ШІ означає моделювання людського інтелекту за допомогою комп'ютерних систем. ШІ пройшов значну еволюцію від перших експертних систем до глибокого навчання, яка відкрила нові можливості для застосування інтелектуальних технологій в усіх сферах життя.

Сьогодні ШІ має прикладне значення у багатьох галузях: у медицині - діагностика за знімками, банківській справі -фінансовий скоринг, транспорті - автопілоти, енергетиці -оптимізація навантажень, обороні -розвідка та моделювання загроз, сільському господарстві -прогнозування врожайності тощо [8].

В українському контексті розвиток ШІ розглядається як стратегічний напрям, у 2020 році Кабінет Міністрів України затвердив **Концепцію розвитку штучного інтелекту в Україні до 2030 року** [11], що передбачає створення нормативно-правової, наукової та інфраструктурної бази для розвитку сфери. Одним з ключових завдань, в яких ШІ відіграє ключову роль є забезпечення безпеки інформаційного простору України та боротьба з дезінформацією.

ШІ дедалі частіше використовується як потужний спосіб для виявлення та протидії дезінформації, його алгоритми здатні аналізувати мільйони текстів на наявність маніпуляцій, фейкових фактів або ворожої пропаганди.

Наприклад, системи з обробки природної мови Natural Language Processing можуть виявляти маніпулятивний стиль, приховані меседжі та порівнювати інформацію з достовірними джерелами.

Технології машинного навчання також застосовуються для створення систем моніторингу соціальних мереж, що автоматично виявляють ворожі інформаційні кампанії — особливо в умовах виборів або гібридної війни, яку веде Російська Федерація проти України.

Разом із тим, використання ШІ породжує і виклики: проблема прозорості алгоритмів (принцип explainable AI), обмеженість навчальних вибірок, ризики упередженості в даних (bias), а також загрози, пов'язані з deepfake-технологіями — системами, що використовують ШІ для створення фейкових відео.

Правове регулювання ШІ є важливою умовою його безпечного розвитку. В Україні відповідне регулювання базується на таких нормативно-правових актах:

- Закон України «Про основні засади забезпечення кібербезпеки України»[9];
- Постанова КМУ про затвердження Концепції розвитку ШІ (2020 рік) [11];
- Адаптація положень європейського **AI Act** – Регламенту ЄС щодо ШІ [10].

SWOT- аналіз розвитку штучного інтелекту в Україні:

Сильні сторони (Strengths):

- Автоматизація процесів. ШІ може значно прискорити процеси збору, аналізу та публікації новин, що дозволяє знижувати витрати і покращувати ефективність.

- Обробка великих даних. ШІ допомагає журналістам швидше обробляти великий обсяг інформації та виявляти важливі патерни чи тренди в даних.

- Персоналізація контенту. Алгоритми можуть допомогти створювати індивідуальні новини для читачів на основі їхніх інтересів і попередньої поведінки.

- Протидія фейковим новинам. Використання ШІ для виявлення фальшивої інформації та автоматичної перевірки може підвищити довіру до медіа.

Слабкі сторони (Weaknesses):

- Недостатнє розуміння штучного інтелекту. В Україні бракує достатньої кваліфікації та знань у журналістів щодо використання ШІ, що може призводити до помилок у роботі.
- Залежність від технологій. Може виникнути надмірна залежність від автоматизованих систем, що зменшує роль людини в журналістиці.
- Етичні питання. В журналістиці використання ШІ може створити питання до етики, зокрема щодо приватності та маніпуляції контентом.
- Висока вартість впровадження. Витрати на впровадження ШІ можуть бути занадто великими для деяких медіа-шкіл стартапів або ін.

Можливості (Opportunities):

- Розвиток нових технологій. Подальший розвиток ШІ відкриває нові можливості для автоматичного створення контенту, таких як генерація текстів, синтезування відео та аудіо.
- Зростання інтересу до даних. На інструменти для аналізу даних у журналістиці в Україні може зрости попит, що забезпечить розвиток цього напрямку.
- Міжнародна співпраця. В галузі штучного інтелекту та журналістики Україна може розвивати співпрацю з міжнародними компаніями та університетами, що сприятиме кардинально новим дослідженням та інноваціям.
- Пристосування до змін у медіапейзажі. Перевагою змін у традиційних медіа може стати поява нових форматів та інструментів для журналістики.

Загрози (Threats):

- Конкуренція з міжнародними ЗМІ. Для українських медіа, що не встигають впроваджувати новітні технології, великі міжнародні медіа-компанії, які активно використовують ШІ, можуть створити конкуренцію.
- Сумнівність регулювання. До правових та етичних проблем може призвести те, що в Україні відсутня чітка правова база щодо використання ШІ у журналістиці.

- Ризик махінацій. Для маніпулювання інформацією або створення дезінформації може бути використано штучний інтелект , що створює ризики для незалежності журналістики.

- Зниження якості контенту. До зниження якості контенту в медіа може призвести надмірне використання автоматизованих систем.

Штучний інтелект стає ключовим інструментом аналізу та захисту цифрового простору, зокрема в умовах кіберзагроз та інформаційної війни. Його розвиток в Україні має відповідати як міжнародним стандартам, так і потребам національної безпеки.

1.3. Огляд наукових підходів та сучасних досліджень у сфері боротьби з дезінформацією

В сучасному інформаційному суспільстві проблема дезінформації стала однією з найбільш актуальних, дезінформація, яка часто включає в себе фейкові новини, маніпулятивні повідомлення, неправдиві або спотворені дані, може мати серйозні наслідки для суспільства, політики, економіки та безпеки [15]. Боротьба з дезінформацією набуває все новіших форм і методів, включаючи використання штучного інтелекту (ШІ), що можна визначити з огляду на стрімкий розвиток цифрових технологій і все більшу роль соціальних мереж у поширенні інформації.

Сучасні дослідження в галузі боротьби з дезінформацією - у цьому підрозділі мною було розглянуто, особливо ті, які орієнтовані на застосування ШІ та автоматизацію процесів виявлення та запобігання поширенню фальшивої інформації.

Проблема дезінформації в медіа

Дезінформація — це свідомо або несвідомо поширена неправдива або спотворена інформація, яка вводить в оману аудиторію і може викликати негативні наслідки для суспільства [16]. У контексті медіа вона може виникати

внаслідок маніпуляцій, вигадування подій або через недбалість у перевірці фактів.

Основні чинники поширення неправдивої інформації-

- швидкість розповсюдження через онлайн-платформи;
- анонімність та відсутність перевірки джерел;
- емоційний вплив та сенсаційність;
- цілеспрямовані інформаційні кампанії та пропаганда [17].

Технології ШІ у боротьбі з дезінформацією

В умовах постійного інформаційного потоку необхідні інструменти, які можуть ефективно виявляти та зупиняти поширення фейків. ШІ відіграє тут ключову роль:

• **Машинне навчання** — дозволяє виявляти зразки дезінформації. Наприклад, дослідження Zhou et al. (2020) доводять ефективність та якість глибоких нейронних мереж у розпізнаванні фейкових новин, використовуючи техніки зору для оцінки змісту новинних статей і визначення маніпуляцій. У роботах Li et al. (2021) також аналізується застосування алгоритмів машинного навчання для класифікації фейкових новин, зокрема, через виявлення семантичних особливостей[18].

• **Обробка природної мови (NLP)** — Д ля виявлення маніпуляцій і неправдивих тверджень у текстовій інформації технології обробки природної мови стали важливим інструментом. Так, Graves (2018) представив методи семантичного аналізу для виявлення контекстуальних маніпуляцій і заперечень фактів на основі лінгвістичних структур і семантики повідомлень [19].

Дослідження також продемонстрували, як NLP можна використовувати для автоматичного виявлення спотворених або маніпулятивних тверджень у новинах. Автоматизована перевірка фактів — системи використовують бази даних і алгоритми для миттєвого виявлення неправдивих тверджень. Вони застосовують методи машинного навчання для автоматичного порівняння тверджень із

відомими фактами.

Аналіз поведінки— інструменти аналізують патерни поширення повідомлень і виявляють організовані кампанії через соціальні мережі.

Також вони досліджують, як бот-активність і інші автоматизовані механізми впливають на дистрибуцію фейкових новин, та як можна виявляти ці патерни для запобігання їхнього поширення[19].

Сучасні дослідження

Окремі проєкти та ініціативи формують академічне підґрунтя боротьби з фейками:

- ClaimBuster та Fake News Challenge — ці проєкти зосереджені на автоматизації перевірки фактів, а також розробці нових алгоритмів для визначення достовірності інформації в реальному часі. Керівники цих проєктів, зокрема, Wang et al. (2017) та Ruchansky et al. (2017), зробили значний внесок у розвиток методів глибинного навчання для детекції фейкових новин [20].
- В роботах Wang (2017) та Ruchansky et al. (2017) аналізується застосування глибинного навчання до виявлення неправдивих новин. Дослідження показують, як ці методи дозволяють поліпшити точність класифікації та зменшити рівень помилкових спрацьовувань при виявленні фейкових новин.
- Інструменти такі як PHEME Project інтегрують соціальні сигнали та ШІ для аналізу достовірності новин в Інтернеті.

Дослідження, проведене командою з PHEME, показало, як соціальні сигнали (поширення через певні мережі, коментарі та реакції користувачів) можуть бути корисними для виявлення та оцінки фейкових новин

Міжнародні ініціативи:

Організації, як EU DisinfoLab, UNESCO та Google News Initiative активно розробляють технології й освітні програми проти дезінформації, підкреслюючи роль співпраці між державами академічною спільнотою й IT-компаніями [21].

Виклики

• **Технічні**: алгоритми не завжди розуміють контекст або сатиру;
• **Етичні**: існує ризик надмірного цензурування, що може порушувати свободу вираження поглядів та доступ до інформації;

• **Освітні**: ключовим лишається розвиток критичного мислення та медіаграмотності серед громадян для ефективної боротьби з дезінформацією

Нові наукові підходи та тренди в дослідженнях

Останні роки характеризуються активним розвитком міждисциплінарних досліджень у сфері боротьби з дезінформацією що охоплюють інформатику, психологію, лінгвістику, соціологію та журналістику. Дослідники шукають не лише технічні рішення, а й соціально-когнітивні механізми, які пояснюють, чому фейкова інформація поширюється і чому в неї вірять.

• **Когнітивні упередження та емоційний вплив** — дослідження, як-от Pennycook & Rand (2019), демонструють, що користувачі часто поширюють дезінформацію не через брак знань, а через когнітивні спрощення — наприклад, ефект підтвердження (confirmation bias), коли люди схильні вірити в те, що відповідає їхнім поглядам [22]. Вивчення таких упереджень допомагає розробляти кращі інтерфейси та повідомлення, які сприяють зупиненню поширення фейків.

• **Фактчекінг як інтерактивний процес** - нові підходи трактують перевірку фактів не просто як технічну дію, а як соціальну взаємодію. Проекти, як-от MediaWise чи PolitiFact, створюють освітні платформи для молоді та дорослих, що формують звички критичного споживання новин.

• **Детекція зображень та відео-фейків** - завдяки вдосконаленню технологій, дезінформація все частіше включає візуальний компонент. Дослідники, як Verdoliva аналізують використання convolutional neural networks (CNN) для виявлення глибоких підробок у відео та зображеннях, які важко відрізнити від справжніх [23].

• **Мультимодальні системи виявлення** — сучасні системи, такі як MultiModalDetector (Chen et al., 2021), комбінують текстовий, візуальний і

метадані-контент для більш точної оцінки достовірності інформації, особливо у випадку соціальних мереж, де пости мають складну структуру.

Український контекст

В Україні проблема дезінформації набула критичного значення через гібридну війну з боку Російської Федерації. Дослідження, проведені Центром стратегічних комунікацій та Інститутом масової інформації (IMI), вказують на високий рівень організованих інформаційних атак. Вітчизняні проєкти, як-от **StopFake**, **Texty.org.ua**, **По той бік новин**, стали прикладом успішного поєднання журналістських практик із елементами аналітики даних.

- StopFake, створений ще у 2014 році, не лише верифікує факти, а й аналізує риторику ворожої пропаганди, використовуючи елементи дискурс-аналізу.
- Texty.org.ua розробили аналітичні інструменти для візуалізації розповсюдження фейків і створення карт інформаційних впливів, які дозволяють ідентифікувати джерела та логіку поширення маніпуляцій.
- Дослідники з Національного університету «Києво-Могилянська академія» вивчають вплив інформаційних вкидів на формування суспільної думки в Україні, комбінуючи методи лінгвістики, соціології та ШІ.

Здатність розрізняти правду від маніпуляцій стає критично важливою навичкою у сучасному світі, де інформація поширюється з блискавичною швидкістю. У цьому розділі я окреслила складну природу дезінформації та простежила, як вона впливає на суспільні процеси, викликаючи довготривалі соціальні, політичні й навіть безпекові наслідки.

Також було розглянуто, як стрімкий розвиток штучного інтелекту відкриває нові горизонти для боротьби з цим викликом, пропонуючи як технічні рішення, так і інструменти для підтримки інформаційної гігієни. Сучасні дослідження і приклади реальних проєктів демонструють, що ефективна протидія дезінформації можлива лише за умов поєднання інноваційних технологій, міждисциплінарного наукового підходу та активної громадянської позиції. Цей

розділ створює фундамент для подальшого глибшого аналізу — зокрема, практичних механізмів виявлення фейкових новин і тих рішень, які вже сьогодні змінюють інформаційний простір.

РОЗДІЛ 2. АНАЛІЗ ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В ЗАРУБІЖНИХ ТА ВІТЧИЗНЯНИХ МЕДІА ДЛЯ ПРОТИДІЇ ДЕЗІНФОРМАЦІЇ

2.1. Методи та технології штучного інтелекту в розпізнаванні фейкових новин

Інформаційна безпека сучасного суспільства дедалі більше залежить від здатності швидко й ефективно знаходити та протидіяти поширенню фейкових новин. З розвитком цифрових соціальних мереж, маштаби дезінформації значно зросли. Згідно з дослідженням MIT, фейкові новини поширюються у шість разів швидше, ніж правдиві [29]. У відповідь на ці виклики зростає роль штучного інтелекту як інструменту виявлення маніпуляцій, фальсифікацій та інформаційних атак.

Машинне навчання - основа автоматизованого розпізнавання.

Машинне навчання є фундаментом більшості інструментів штучного інтелекту в аналізі нового контенту. Алгоритми машинного навчання використовуються для побудови класифікаторів, які здатні відрізнити фейкову новину від правдивої на основі набору ознак. До таких ознак належать:

- частота використання сенсаційної лексики;
- надмірна кількість прикметників або емоційно забарвлених слів;
- наявність заголовків у стилі "клікбейт";
- нестача джерел, або відсутність посилання на офіційні документи [30].

Зокрема, демонструють високу точність у задачах двокласової класифікації моделі, побудовані на основі Logistic Regression або Naive Bayes. У дослідженнях [31] вказується, що за умови належної підготовки даних точність таких моделей може сягати понад 85%.

Глибинне навчання та трансформери.

З появою глибинного навчання (Deep Learning) якість виявлення фейкових новин значно покращилась.

Архітектури LSTM (Long Short-Term Memory) дозволяють враховувати контекст речень, що особливо корисно при проведенні аналізу довгих текстів новин.

Однак, справжній прорив забезпечили моделі на базі трансформерів, зокрема:

- BERT (Bidirectional Encoder Representations from Transformers);
- RoBERTa;
- GPT (Generative Pre-trained Transformer).

Ці моделі враховують контекст слова у реченні, розуміють зв'язки між словами та можуть генерувати логічні висновки. Дослідження [32] показують, що у задачах з фактчекінгу трансформери досягають точності понад 90%, якщо їх навчено на достатньо великій базі данних.

Обробка природної мови (NLP) як головний інструмент.

Інструменти обробки справжньої мови, дозволяють не лише класифікувати новини, а й аналізувати їх зміст, логічні зв'язки, їх стилістику та джерела.

Основні методи NLP, що використовуються:

- аналіз сентименту — визначає емоційне забарвлення тексту;
- виявлення іменованих сутностей (NER) — допомагає ідентифікувати згаданих осіб, організації та місця;
- виявлення протиріч — визначає, чи суперечить твердження раніше встановленим фактам [33].

Інструменти обробки природньої мови стали за основу для створення автоматизованих систем фактчекінгу, таких як ClaimBuster, які здатні у реальному часі відзначати потенційно підозрілі твердження у промовах політиків або публікаціях ЗМІ [34].

Гібридні системи: об'єднання методів для максимального ефекту.

Успішне розпізнавання фейкових новин зазвичай не обмежується застосуванням одного алгоритму. Найефективнішими вважаються **гібридні системи**, які поєднують:

- аналіз метаданих (дата, джерело, авторство);
- перевірку фактів через бази знань (Google Fact Check Tools, Wikidata);
- аналіз стилістичних та семантичних особливостей тексту [35].

Такі системи мають змогу працювати в реальному часі, попереджаючи користувачів про потенційно небезпечний контент.

Поведінковий аналіз та мережеві патерни

Важливим аспектом роботи ШІ у сфері боротьби з дезінформацією є аналіз розповсюдження інформації. Дослідження показують, що фейкові новини часто мають специфічні патерни поширення:

- масові репости у перші години після публікації;
- активність бот-акаунтів;
- повторне використання однакових шаблонів повідомлень [36].

Моделі, які аналізують поведінку користувачів (наприклад, Botometer), можуть виявляти аномальні мережеві активності та маркувати їх як підозрілі.

Український контекст: виклики та досягнення

В Україні проблематика фейкових новин особливо актуальна в умовах інформаційної війни, спрямованої на дестабілізацію громадської думки та підрив довіри до державних інституцій. У відповідь на ці виклики були створені ініціативи, зокрема:

- **StopFake** — платформа, яка першою в Україні розпочала перевірку новин на фейковість;
- **VoxCheck** — аналітичний центр, який поєднує людський фактчекінг з технологіями NLP;
- **OSINT-команди при Міноборони** — що використовують автоматизовані алгоритми моніторингу російських інформаційних атак [37].

Однак, як зазначено у звіті Ради національної безпеки і оборони України [38], інтеграція штучного інтелекту в державну політику потребує значних інвестицій, міжнародного досвіду та стандартизації процесів.

2.2. Огляд сучасних інструментів та алгоритмів штучного інтелекту для аналізу інформації

У боротьбі з дезінформацією головну роль відіграє здатність швидко, точно та масштабно аналізувати великі обсяги даних.

Завдяки розвитку у сфері штучного інтелекту, почали з'являтися потужні інструменти, що дозволяють не лише знаходити фейкові повідомлення, а й робити масштабний моніторинг простору медіа в режимі реального часу. Цей підрозділ присвячено детальному розгляду сучасних алгоритмів і рішень, які використовуються в галузі аналізу інформації з метою протидії дезінформації.

Алгоритми класифікації - це основа боротьби з фейками. Алгоритми класифікації — це основа багатьох інструментів аналізу контенту. Найчастіше використовують такі моделі:

• Naïve Bayes — проста, але ефективна модель, що бсформована на теоремі Байєса. Особливо гарно працює з короткими текстами, наприклад, твітами [39].

• Support Vector Machines (SVM) — дозволяє розділяти великі обсяги даних за допомогою гіперплощин; часто застосовується в задачах розпізнавання емоційного забарвлення тексту, або правдивості чи фейковості новин [40].

• Logistic Regression — класичний статистичний підхід, який добре та стрімко розвивається та інтерпретується.

Ці моделі використовуються в таких відомих системах, як LIAR dataset classifier або FakeFinder, які дозволяють поділяти повідомлення за рівнем правдивості [41].

Нейромережеві архітектури: від CNN до GPT

Нові етапи розвитку штучного інтелекту, пов’язані з появою глибинних моделей аналізу інформації -

• CNN (Convolutional Neural Networks) — хоча спочатку створені для зображень, вони адаптовані для аналізу тексту, особливо для виявлення особливостей фейкових новин [42].

• LSTM (Long Short-Term Memory) — ефективні при аналізі послідовностей слів, можуть запам'ятовувати попередні частини тексту.

• Transformers (BERT, RoBERTa, XLNet) — моделі які сьогодні є лідерами у сфері обробки природної мови, вони розуміють контекст на рівні абзаців, а не лише просто ключових слів [43].

• GPT (Generative Pre-trained Transformer) — використовується не лише для створення тексту, а й для його аналізу та виявлення фактологічних помилок, маніпуляцій, неетичних висловлювань тощо.

Цікавий факт: Модель BERT, розроблена Google у 2018 році, настільки ефективно зрозуміла контекст, що її інтегрували у власну пошукову систему, зробивши її точнішою для мільярдів запитів щодня [44].

Інструменти перевірки фактів на основі ШІ.
У глобальній практиці сьогодні активно використовуються повністю автоматизовані системи фактчекінгу, серед найпоширеніших:
• ClaimBuster (США) — виявляє перевіркові твердження у політичних дебатах, або статтях та порівнює їх із базами достовірних джерел [45].
• Full Fact (Велика Британія) — автоматизований фактчекер, що поєднується з парламентськими дебатами в режимі реального часу.
• FABULA AI — британський стартап, який застосовує графовий аналіз для поширення новин у соцмережах, а не лише аналіз змісту новин. Він розпізнає фейки за тим, як вони подорожують онлайн-простором [46].

NLP-інструменти для виявлення маніпуляцій і пропаганди
Обробка природної мови — ключ до розпізнавання контексту прихованої агресії, що часто використовуються у фейках.
Серед актуальних технологій в першу чергу це-
• Named Entity Recognition (NER) — виділення осіб, подій, організацій;
• Sentiment Analysis — визначення тональності тексту (емоційне забарвлення);
• Discourse Analysis — розуміння логіки побудови тверджень, виявлення маніпулятивних переходів;
• Stance Detection — аналіз відношення автора до теми (підтримка, заперечення тощо).

Цікавий факт: У 2022 році дослідники з Массачусетського технологічного інституту довели, що тексти з фейкових новин мають набагато більше емоційне забарвлення, більше використовують крайні судження та рідше містять посилання на першоджерела інформації [47].
Гібридні підходи – це інтеграція ШІ з людською експертизою. Найефективніші інструменти — це ті, що поєднують штучний інтелект з участю людини. Наприклад Media Bias/Fact Check (MBFC) застосовує автоматизований аналіз для попереднього оцінювання джерела, після чого модератори перевіряють вручну, а The Washington Post Fact Checker AI працює у форматі редактор та алгоритм, ШІ аналізує а людина верифікує.
Цей гібридний підхід є особливо актуальним для країн, де ШІ-системи ще не досягли повної автоматизованості у складних політичних або культурних ситуаціях.

Перспективні напрямки розвитку. Explainable AI (XAI) — ШІ, який може пояснити, чому він вважає інформацію фейковою. Це критично важливо для довіри користувачів.
Мультимодальні моделі здатні одночасно аналізувати текст, зображення, відео, аудіо, такі підходи стануть головними у виявленні deepfake - контенту.
Інтеграція з blockchain — дозволяє створити цифрові сліди правдивої інформації, запобігаючи її підробці в майбутньому [48].
Сучасні інструменти штучного інтелекту для аналізу інформації показують значний розвиток у виявленні дезінформації, однак жоден із них не є універсальним для будь- чого. Найкраща ефективність досягається за допомогою поєднання кількох методів: від аналізу структури повідомлення до знаходження аномалій у поширенні та змісті. Перспективним є розвиток пояснюваного ШІ, мультимодальних систем і моделей, які адаптуються до локального контексту.

2.3. Дослідження практик використання ШІ в зарубіжних та вітчизняних медіа

Із розвитком цифрових технологій та зростанням впливу соціальних мереж на формування громадської думки, мас-медіа дедалі активніше впроваджують штучний інтелект (ШІ) для автоматизації процесів створення, поширення й перевірки контенту. У цьому підрозділі розглядаються практики використання ШІ в зарубіжних і вітчизняних медіа, зокрема у боротьбі з дезінформацією, з прикладами конкретних рішень і платформ.

1. Зарубіжні приклади використання ШІ в медіа

1.1. The Washington Post: Heliograf.

Ще у 2016 році американське видання The Washington Post запустило Heliograf — автоматизовану платформу на основі ШІ для створення коротких новин. Під час Олімпійських ігор та виборів вона згенерувала понад 850 новин. Особливість платформи — інтеграція з аналітичними модулями, які дозволяли виявляти нестандартні дані та аномалії, що могли свідчити про дезінформацію [58].

1.2. BBC: модуль інтелектуального фактчекінгу.

BBC Reality Check використовує ШІ-модулі для моніторингу заяв політиків і публікацій у соцмережах. Алгоритми виділяють ключові твердження, які потім проходять ручну перевірку. Система також взаємодіє з BERT-моделями для попереднього аналізу фактологічних помилок [59].

1.3. Reuters та алгоритми Trust Principles.

Інформагентство Reuters інтегрувало ШІ-модулі для оцінки "надійності джерел" відповідно до принципів Trust Principles. Система аналізує походження інформації, рівень цитованості та наявність упередженості [60].

1.4. Meta AI: RoBERTa for Fake News.

Meta розробила модель "RoBERTa for Fake News", яка аналізує мільйони постів щодня, ідентифікуючи фейки, сатиру, упереджені публікації та спонсоровану пропаганду [61].

1.5. The Guardian: RADAR AI.

У співпраці з Press Association, The Guardian застосовує систему RADAR — автоматизовану ШІ-платформу для створення локальних новин. RADAR генерує тексти на основі відкритих даних, а редактори перевіряють та допрацьовують матеріал. У 2019 році було опубліковано понад 100 тис. таких новин [62].

1.6. New York Times: Perspective API.

New York Times тестує систему від Jigsaw (підрозділ Google) — Perspective API. Цей ШІ-інструмент аналізує токсичність коментарів під статтями й автоматично приховує маніпулятивні або емоційно агресивні повідомлення, що є частиною боротьби з дезінформаційною агресією в коментарях [63].

2. Вітчизняні практики використання ШІ

2.1. Суспільне та фактчекінгові ініціативи.

Суспільне застосовує напівавтоматизовані фільтри новин із використанням українськомовної BERT-моделі (UBERT), що дозволяє на ранньому етапі виявляти маніпуляції, фейки та спотворення контексту [64].

2.2. VoxCheck AI.

Інструмент VoxCheck AI автоматично виявляє маніпулятивні висловлювання публічних осіб та звіряє їх із наявними фактчекінговими базами даних. Планується інтеграція з платформами відеоконтенту [65].

2.3. Detector Media.

Detector Media впровадив ШІ-систему для моніторингу дезінформації, виявлення бот-активності та аналізу координації інформаційних атак [66].

2.4. Бот «Перевірка новин» від Мінцифри.

Бот створено у партнерстві з організаціями «Інтерньюз» і «Медіасапієнс». Він аналізує новини, оцінює достовірність та вчить користувачів перевіряти інформацію самостійно [67].

2.5. Ініціатива “ПравдаBot” від StopFake.

Бот використовує NLP-алгоритми для сканування соціальних мереж і месенджерів, аби виявити повторювані фейки, які поширюються організовано.

Його функції включають попередження користувача та пропозицію достовірного джерела [68].

3. Міжнародні та міжсекторальні ініціативи

- AI4Media.
  Європейський проєкт, до якого залучені українські медіаексперти, має на меті створити інноваційні та етичні стандарти використання ШІ в медіа [69].
- Google News Initiative.
  Ця платформа підтримує українські медіа у впровадженні ШІ-рішень для перевірки фактів, боротьби з фейками та виявлення маніпуляцій [70].
- UNESCO AI and Journalism Ethics.
  Інструмент навчання журналістів етичному використанню ШІ, з акцентом на цифрову безпеку, розпізнавання фейків і прозорість алгоритмів [71].

Досвід зарубіжних та українських медіа підтверджує, що ШІ стає ключовим інструментом у боротьбі з дезінформацією. Із автоматизації новин до аналізу джерел та виявлення фейків — штучний інтелект трансформує підходи до журналістики. Україна активно долучається до глобальних трендів, і попри війну, показує інноваційні підходи до інформаційної безпеки.

2.4. SWOT-аналіз ефективності використання штучного інтелекту в боротьбі з дезінформацією

SWOT-аналіз — це інструмент стратегічного планування, який дозволяє оцінити сильні сторони (Strengths), слабкі сторони (Weaknesses), можливості (Opportunities) та загрози (Threats) певної технології або підходу. У випадку штучного інтелекту (ШІ) у сфері боротьби з дезінформацією, SWOT-аналіз допомагає зрозуміти як потенціал, так і ризики його впровадження у медіапрактику.

Сильні сторони(Strengths)

Інтеграція ШІ в засоби спілкування для перевірки фактів

У 2025 році платформа Perplexity ввела інструмент перевірки фактів у WhatsApp, що дозволяє користувачам надсилати повідомлення для миттєвої перевірки правильності інформації. Цей інструмент підтримує понад 20 мов, що робить його ефективним у багатомовних спільнотах, де дезінформація може швидко розповсюджуватися через групові месенджери.

Розробка спеціалізованих інструментів для виявлення дезінформації

Одна з розробок компанії AI Seer у Сінгапурі, створила новий інструмент перевірки фактів, який використовує штучний інтелект для аналізування текстових та відеоматеріалів. Цей інструмент досягає точності 92% у виявленні фейкових новин, що перевищує показники інших платформ.

Використання ШІ в для боротьби з дезінформацією

Україна активно впроваджує ШІ для протидії російській дезінформації, наприклад, бот "Перевірка", розроблений Центром протидії дезінформації при РНБО, дозволяє громадянам надсилати підозрілі повідомлення для перевірки, отримуючи відповіді через декілька хвилин .

Швидка обробка великих обсягів даних

Алгоритми ШІ можуть аналізувати тисячі джерел інформації одночасно, наприклад, платформа Blackbird відстежує інформаційні атаки в реальному часі, автоматично виявляючи джерела фейків, їхні зв’язки та шляхи поширення [1].

Автоматичне розпізнавання фейків

Моделі штучного інтелекту використовуються в таких системах, як ClaimBuster (США), яка автоматично позначає потенційно неправдиві заяви політиків та відправляє їх на перевірку аналітикам [2]. Аналогічно, в ЄС використовують Truly Media — платформу, що поєднує штучний інтелект та краудсорсинг для перевірки контенту.

Нейтральність алгоритмів у прийнятті рішень

На відміну від людей, ШІ не має політичної чи емоційної упередженості, що дає йому можливість затверджувати рішення на основі даних.

Наприклад, модель BERT показала хороші результати у класифікації фейків завдяки аналізу семантики без впливу упереджень [3].

Прогнозування інформаційних атак

Інструменти на кшталт ThreatMapper аналізують попередні патерни дезінформації та можуть заздалегідь попередити про нові заходи, що набирають обертів. Вони фіксують різке зростання активності ботів, ідентичність повідомлень та їх поширення [4].

Скорочення людських витрат

Завдяки автоматизації процесів перевірки, редакції новин можуть скоротити витрати на аналітику, наприклад, у Reuters ШІ допомагає журналістам швидко перевіряти джерела[5].

Слабкі сторони (Weaknesses)

Чутливість до маніпуляцій

Зловмисники використовують генеративний ШІ для створення глибоких фейків, які важко відрізнити від реальної інформації. Це ускладнює завдання ШІ-системам, які мають виявляти такі фейки, особливо коли вони створені з високою якістю

Обмеження в розпізнаванні мультимодальної дезінформації

Більшість існуючих моделей ШІ орієнтовані на текстовий контент і мають труднощі з аналізом мультимедійних матеріалів, які поєднують текст, зображення та відео. Це обмежує їхню ефективність у виявленні комплексної дезінформації. ШІ іноді не відрізняє іронію від справжньої інформації. Наприклад, с видання "The Onion" неодноразово визнавалося фейковим деякими алгоритмами детекції [6].

Обмежена підтримка мов та локального контенту

Більшість алгоритмів розробляються англійською. Для української мови досі бракує потужних моделей. Наприклад, StopFake.org використовує ШІ, але багато фейків залишаються поза зоною дії [7].

Ризик автоматичних помилок і "перекосів"

Якщо дані, на яких навчався ШІ, містять упередження, алгоритм буде їх відтворювати. Це актуально у випадках, коли інформація з певних джерел автоматично вважається неправдивою без контексту. Наприклад, деякі моделі вважали повідомлення з джерел у Східній Європі менш надійними через географічний фактор [8].

Закритість алгоритмів і відсутність пояснень

Чорна скринька моделей (особливо трансформерів) означає, що часто неможливо пояснити, чому певна новина була позначена як фейкова. Це створює проблеми довіри до систем і потребує впровадження Explainable [9].

Недостатня інтеграція в медіапрактики

Багато редакцій не мають спеціалістів з ШІ або технічної інфраструктури для роботи з такими інструментами. Наприклад, у більшості регіональних ЗМІ штучний інтелект або не використовується взагалі, або застосовується у вигляді простих пошукових фільтрів.

Можливості (Opportunities)

Розвиток прозорих стандартів та етики

Ініціативи на зразок EU AI Act та UNESCO Recommendation on AI Ethics формують глобальні рамки для етичного використання ШІ. Це відкриває можливості для створення справедливих систем перевірки контенту [10].

Освітні платформи з підтримкою ШІ

Chat-боти з елементами NLP використовуються для тренування критичного мислення в школах. Наприклад, в Іспанії у 2023 році запущено проєкт FaktualBot, який навчає дітей розпізнавати фейки через гейміфікацію [11].

Публічні API для ЗМІ

Все більше інструментів відкривають свої API (наприклад, Google FactCheck Tools API, Media Cloud), що дозволяє ЗМІ інтегрувати ШІ у власні редакційні системи [12].

Національні центри протидії дезінформації з використанням ШІ

В Україні вже створено Центр стратегічних комунікацій при Мінкульті. Якщо інтегрувати в нього ШІ-модулі для аналізу соцмереж та новин, це значно підвищить ефективність реагування [13].

Глобальна співпраця між урядами, ІТ та наукою

Прикладом є ініціатива Partnership on AI, яка об'єднує такі компанії, як Google, Facebook, Microsoft, а також наукові інституції — для спільного вдосконалення технологій перевірки фактів.

Загрози

Використання ШІ для створення дезінформації

Технології deepfake та generative AI (напр. GPT, Claude) можуть створювати фейки, що візуально та текстуально не відрізняються від справжніх новин. У 2023 році в Індії було виявлено deepfake-відео політика, яке поширювалося в TikTok і отримало мільйони переглядів за годину [14].

Зловживання владою

Деякі авторитарні режими використовують системи ШІ для контролю за медіа, блокування "незручних" матеріалів або цензури. У Китаї, наприклад, використовують ШІ для автоматичного фільтрування "шкідливого контенту", що часто включає й правдиві новини [15].

Підрив довіри до журналістики

Якщо алгоритм помиляється, ідентифікуючи правдиву новину як фейк, аудиторія втрачає довіру як до ЗМІ, так і до інструментів перевірки. Це вже відбувалося у США, коли Facebook блокував новини, які потім виявилися правдивими, через помилки алгоритмів[16].

Кіберзагрози та атаки на моделі

Системи ШІ можна атакувати — змінюючи вхідні дані так, щоб модель видавала неправильні результати. Це називається "adversarial attacks". У 2022 році дослідники продемонстрували, як фейковий акаунт може зламати систему бот-детекції, змінюючи лише кілька слів у повідомленнях [17].

Штучний інтелект має значний потенціал у протидії дезінформації, однак потребує відповідального підходу, етичного контролю, локалізації та пояснюваності. SWOT-аналіз демонструє, що ефективність ШІ залежить не лише від його технічної досконалості, а й від соціального контексту, в якому його застосовують.

У цьому розділі було детально розглянуто використання штучного інтелекту (ШІ) в боротьбі з дезінформацією, зокрема у сфері медіа. Аналіз методів та технологій ШІ, застосовуваних для розпізнавання фейкових новин, дозволяє відзначити високий потенціал цих інструментів у швидкому та ефективному виявленні маніпуляцій з інформацією. Завдяки використанню таких технологій, як глибинне навчання, машинне навчання, алгоритми на основі природної мови (NLP) та нейронні мережі, медіа можуть значно підвищити точність перевірки фактів, автоматизувати процеси та знижувати людські помилки в аналізі новин.

Огляд сучасних інструментів та алгоритмів, таких як ClaimBuster, Truly Media, Blackbird.AI, а також інтеграція новітніх моделей штучного інтелекту, демонструє значні досягнення в розпізнаванні дезінформації на різних рівнях – від автоматичного виявлення фейкових новин до прогнозування можливих інформаційних атак. Крім того, впровадження ШІ в медіапрактики дозволяє створювати інструменти для більш ефективної боротьби з інформаційними маніпуляціями, що має особливо велике значення в умовах глобалізації та зростаючого впливу дезінформаційних кампаній.

Проте, попри переваги, застосування штучного інтелекту в цій сфері супроводжується певними викликами.

Обмеження, пов'язані з контекстом, мовною підтримкою, ризиком помилок та упереджень в алгоритмах, вимагають подальших досліджень і вдосконалення інструментів для підвищення їх ефективності. Важливою є також етична складова, що вимагає прозорості та пояснюваності алгоритмів для забезпечення довіри з боку суспільства.

Загалом, використання ШІ в протидії дезінформації представляє собою потужний інструмент для медіа, однак для забезпечення його максимального ефекту потрібно активно враховувати соціальний контекст, технічні обмеження та етичні норми. Тому подальше вдосконалення технологій та їх інтеграція в медіапрактики потребують комплексного підходу, що включає міждисциплінарне співробітництво між науковцями, розробниками та медіа-професіоналами.

РОЗДІЛ 3.
РЕКОМЕНДАЦІЇ ТА ПРОПОЗИЦІЇ ЩОДО ВДОСКОНАЛЕННЯ ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ У ПРОТИДІЇ ДЕЗІНФОРМАЦІЇ

3.1. Оптимізація методів машинного навчання для виявлення дезінформації

У сучасному світі, де інформація поширюється миттєво завдяки цифровим технологіям, проблема дезінформації набуває величезного масштабу. В особливості, під час гібридних війн, як це ми спостерігаємо в Україні, фейкові новини використовуються як інструмент психологічної та інформаційної війни. Дезінформація може мати катастрофічні наслідки, зокрема посіяти паніку, дестабілізувати суспільство та змінити громадську думку. Тому розробка та оптимізація методів машинного навчання (МН) для виявлення фейкових новин і дезінформації є критично важливою задачею для сучасних технологій [80].

Машинне навчання дозволяє автоматизувати процеси аналізу великих обсягів даних, що в свою чергу дає змогу виявляти приховані закономірності, які можуть бути незрозумілими або важкодоступними для традиційних методів аналізу. На основі цього, ШІ здатен розпізнавати не лише явні маніпуляції, але й приховані елементи в текстах, зображеннях та відео, які складають дезінформаційні кампанії. Наприклад, сучасні глибокі нейронні мережі, такі як BERT, GPT-3, та трансформери, вже довели свою ефективність у розпізнаванні контексту, сарказму, маніпуляцій і навіть емоційного забарвлення [81].

Алгоритми машинного навчання

Існують різні підходи до машинного навчання для боротьби з дезінформацією. Один із них базується на використанні класичних алгоритмів, таких як логістична регресія, дерева рішень або методи опорних векторів. Однак ці методи мають обмеження, адже вони не здатні обробляти складні семантичні зв'язки в текстах.

Водночас, більш складні моделі глибинного навчання (deep learning), зокрема рекурентні нейронні мережі (RNN) і трансформери, які стали основою для таких технологій, як BERT та GPT, значно покращують точність виявлення дезінформації, завдяки здатності глибше розуміти контекст і знаходити неочевидні патерни в даних [82].

Мультимодальні підходи

Одним із важливих напрямків у боротьбі з дезінформацією є мультимодальні моделі, які працюють не лише з текстовими даними, але й з іншими видами інформації, такими як зображення або відео. Наприклад, аналіз відео для виявлення deepfake або маніпуляцій у зображеннях активно використовується в популярних платформах, таких як YouTube та Facebook. Оскільки такі медіа можуть містити деталі, які важко виявити за допомогою лише текстового аналізу, мультимодальні підходи дозволяють поєднувати текстовий контент із візуальними або аудіо елементами для досягнення більш точного виявлення маніпуляцій [82].

Застосування таких моделей у реальних умовах показало високі результати. Наприклад, у 2020 році одна з платформ, яка займається виявленням фейкових відео, досягла точності 90% у виявленні маніпуляційних відео за допомогою штучного інтелекту. Такі результати відкривають нові можливості для боротьби з маніпуляціями в медіапросторі, де відео та зображення можуть мати великий вплив на громадську думку [83].

Адаптивність та самонавчання

Важливою характеристикою сучасних моделей машинного навчання є їх здатність до самооновлення та адаптації до нових даних. Дезінформація, як і самі алгоритми її поширення, постійно еволюціонує. Зміна формату фейкових новин, використання нових платформ або зміна стилю маніпуляцій потребує того, щоб моделі були здатні не лише реагувати на наявні патерни, але й навчатися від

нових прикладів. Наприклад, через регулярне перенавчання на нових трендових даних, моделі зможуть вчасно виявляти нові дезінформаційні хвилі [80].

Це є надзвичайно важливою складовою, адже сучасні інформаційні війни вимагають від аналітичних систем не лише точності, а й швидкості реагування на нові маніпуляції, що постійно з'являються в медіапросторі. Оскільки дезінформація може з’являтися в різних формах і на різних платформах, необхідно постійно оновлювати моделі, щоб вони залишались актуальними [82].

Виявлення дезінформації в українському контексті

В Україні особливо важливим є врахування мовних, стилістичних і семантичних особливостей текстів, які використовуються в пропагандистських кампаніях. Мова ворожої пропаганди має специфічні ознаки: перебільшення, використання драматизму, маніпуляції з фактами, що часто не відповідають реальності. Проте такі тексти часто маскуються під правдоподібні повідомлення, що ускладнює їх виявлення без застосування спеціалізованих моделей [84].

Додатково, в умовах гібридної війни, де ворожі медіа активізуються не лише через основні новинні платформи, але й через соціальні мережі, важливо розробляти моделі, які здатні виявляти фейки у нових форматах, таких як «меми», відео, або навіть анонімні повідомлення в месенджерах. Тому Україна також активно працює над створенням локалізованих моделей, що враховують не тільки лінгвістичні особливості, а й соціокультурні контексти [85].

Міждисциплінарний підхід

Оптимізація методів машинного навчання для виявлення дезінформації потребує тісної співпраці з іншими галузями науки, такими як лінгвістика, соціологія, психологія та журналістика. Оскільки дезінформація — це не просто набір неправдивих фактів, а й інструмент психологічного впливу, важливо враховувати не лише технічні аспекти, а й людський фактор. Для створення ефективних систем, які зможуть на реальному рівні боротися з фейковими новинами, необхідно залучати фахівців, які розуміють соціальні процеси і можуть

враховувати людські психологічні особливості при розпізнаванні маніпуляцій [80], [81].

3.2. Використання штучного інтелекту в державних і приватних ініціативах боротьби з фейками

У сучасному інформаційному середовищі, де фейкові новини можуть поширюватися зі швидкістю кількох кліків, використання штучного інтелекту (ШІ) стає критично важливим інструментом для захисту суспільства від маніпуляцій. Державні та приватні ініціативи, які інтегрують інструменти ШІ у боротьбу з дезінформацією, відіграють все більш помітну роль у забезпеченні інформаційної безпеки як в Україні, так і на глобальному рівні.

Державний сектор

Державний сектор активно інвестує в системи на основі ШІ, здатні оперативно виявляти, класифікувати та аналізувати потенційно небезпечний або маніпулятивний контент. Однією з ключових українських ініціатив є діяльність Центру протидії дезінформації при РНБО, створеного у 2021 році. Центр використовує алгоритми машинного навчання для аналізу новин, соціальних мереж і публічних комунікацій. Його завдання — не лише виявити фейкову інформацію, а й прогнозувати напрямки нових інформаційних атак, наприклад, через аналіз повторюваних наративів у медіа [86].

Центр активно використовує семантичний аналіз текстів, що дозволяє ідентифікувати приховані смислові патерни та загрози. Враховуючи швидке поширення маніпуляцій під час повномасштабного вторгнення РФ у 2022 році, ШІ став надзвичайно важливим інструментом для моніторингу таких фейкових вкидів, як повідомлення про "здачу" українських міст або "недовіру до командування ЗСУ", що активно поширювалися через бот-мережі [87].

Одним з найцікавіших прикладів використання ШІ в роботі Центру є їхня співпраця з платформами фактчекінгу та моніторингу медіа, які аналізують соціальні мережі для виявлення потенційних загроз. Зокрема, для швидкого

відстеження великомасштабних інформаційних атак було використано технології розпізнавання аномалій, які можуть виявляти патерни, що вказують на скоординовану кампанію дезінформації [88].

Такі технології, як інтелектуальні системи розпізнавання аномалій і прогнозування інформаційних хвиль, стали невід'ємною частиною системи безпеки. Наприклад, в умовах війни урядова структура CERT-UA використовує штучний інтелект для виявлення скоординованих кампаній з дезінформації. Зокрема, вони застосовують системи для аналізу бот-мереж і виявлення нетипових хвиль публікацій у соціальних мережах. З кожним роком алгоритми виявляють нові й складніші форми маніпуляцій, зокрема фальшиві новини, що маскуються під реальні події [89].

Глобальні приклади використання ШІ в державних ініціативах

У світі також існують численні приклади використання ШІ державними структурами для боротьби з фейками. Одним із таких прикладів є платформа EUvsDisinfo, створена Європейським Союзом для моніторингу і боротьби з російською дезінформацією. Платформа використовує потужні алгоритми ШІ, здатні автоматично аналізувати теми, ключові слова та джерела інформації для виявлення пропаганди та фейкових новин. Одна з основних функцій платформи — виявлення фейкових новин, що мають тенденцію до поширення через певні канали, наприклад, через певні пропагандистські ЗМІ, що працюють під контролем державних структур [90].

Особливо цікавою є співпраця Європейського Союзу з іншими міжнародними організаціями та медіа-компаніями, які використовують спільні алгоритми для аналізу контенту на декількох мовах. Завдяки цій співпраці платформа здатна відстежувати не лише фейкові новини, але й інформаційні патерни, характерні для конкретних країн, що дозволяє виявляти нові напрямки вкидів дезінформації, специфічні для кожної країни чи культури [91].

Важливою складовою ефективної боротьби з дезінформацією є також використання ШІ в рамках програм, які працюють з медіа та новими технологіями. Наприклад, у США програма DARPA's Semantic Forensics

(SemaFor) активно реалізує передові технології для розпізнавання фейкових фото, відео і текстів, в тому числі через використання нейронних мереж.

Програма орієнтована на виявлення не лише фальшивих новин, а й розкриття їхніх авторів, а також мотивів кампаній, що їх поширюють. Одним із важливих елементів цієї програми є її здатність автоматично виявляти аномальні зміни в контексті фото або відео, що дозволяє своєчасно виявляти маніпуляції з медіа-контентом, спрямовані на створення фейкових новин або пропаганди [92].

Приватний сектор: інновації у боротьбі з фейками

Приватний сектор також активно розвиває рішення на основі ШІ для боротьби з дезінформацією. Великі технологічні компанії, такі як Facebook (Meta) та Google, використовують інтелектуальні системи для модерації контенту і боротьби з фейковими новинами. Алгоритми Google News, наприклад, аналізують джерела новин на відповідність критеріям достовірності та історії фейків, що допомагає автоматично виявляти потенційно небезпечний контент. Вони працюють за допомогою технологій, що дозволяють оперативно виявляти новини, які можуть бути частиною скоординованих кампаній з дезінформації, що має велике значення для моніторингу інформаційного простору в реальному часі [93].

Однак саме компанія Logically вирізняється своїм підходом до використання ШІ в боротьбі з дезінформацією. Зокрема, їхня платформа для реального часу використовує алгоритми, здатні відстежувати та виявляти приховані зв'язки між джерелами, мовними шаблонами та динамікою поширення фейків. Це дозволяє не лише виявляти фейкові новини, а й передбачати їхнє подальше поширення, що надає змогу здійснювати превентивні заходи. Платформа стала особливо популярною під час пандемії COVID-19, коли теорії змови щодо вакцинації та медичних процедур стали популярними. Платформа Logically дозволила оперативно виявляти такі маніпуляції і позначати потенційно шкідливі новини ще до їх широкого поширення [94].

Ще одним важливим прикладом є діяльність стартапу LetsData, що активно працює з українськими медіа-компаніями для виявлення фейкових новин та загроз у цифровому середовищі.

Їхня система виявлення фейків здатна класифікувати загрози, зокрема бот-атаки, панічні вкиди, спроби політичної дестабілізації, що дозволяє оперативно реагувати на інформаційні атаки та забезпечити захист інформаційної безпеки на державному рівні [95].

Взаємодія між державними і приватними ініціативами

Сучасна ситуація вимагає не тільки розвитку інструментів ШІ, але й тісної співпраці між державними і приватними організаціями для ефективної боротьби з фейками. В Україні, наприклад, була налагоджена тісна співпраця між медіа, громадськими організаціями та ІТ-компаніями для створення платформ фактчекінгу, таких як StopFake, По той бік новин, VoxCheck. Ці платформи поєднують можливості автоматичної перевірки інформації за допомогою ШІ з людською експертизою, що дозволяє забезпечити високу точність виявлення фейків. Вони також стали основою для навчання нових моделей ШІ, що дозволяє системам постійно покращувати свої можливості у боротьбі з дезінформацією [96].

На підставі вищезгаданих прикладів і досліджень можна зробити висновок, що інтеграція ШІ у системи боротьби з дезінформацією в державних та приватних ініціативах забезпечує не лише ефективність виявлення фейків, а й швидкість реакції на нові загрози. Водночас, інтеграція людини в цей процес — через експертну перевірку та етичний контроль — є незамінною для досягнення більш високого рівня точності та гнучкості систем.

3.3. Перспективи розвитку штучного інтелекту в інформаційній безпеці

Розвиток інформаційної безпеки в умовах гібридних загроз, кібервійн, глибоких фейків та інформаційного тероризму неможливий без використання штучного інтелекту (ШІ) як одного з основних інструментів превентивного та реактивного впливу. У XXI столітті ШІ перетворюється на центральний компонент цифрової безпеки, забезпечуючи не лише автоматизовану фільтрацію контенту, а й

здатність до стратегічного аналізу інформаційних кампаній, а також розпізнавання елементів когнітивного впливу на суспільство.

Автоматизовані аналітичні системи нового покоління

ШІ-системи наступного покоління в інформаційній безпеці зосереджені не лише на виявленні, але й на активному запобіганні поширенню дезінформації. Наприклад, розробка автономних систем типу "Cognitive Security AI" дає змогу не лише виявити фейковий контент, але й аналізувати його походження, цілі, наративи та ймовірні шляхи подальшого поширення [97].

Такі системи можуть працювати в умовах мультиканальної обробки даних, зокрема аналізуючи публікації в Telegram, TikTok, YouTube, форумах і dark web. Приклад — проєкт "OpenMind" від компанії Logically, який використовує багаторівневу ШІ-аналітику для визначення потенційної небезпеки інформаційних вкидів на ранньому етапі [98].

Перспектива "інформаційних цифрових щитів"

Візією майбутнього є створення так званих «інформаційних цифрових щитів» (Information Shielding AI), що поєднують машинне навчання, розподілену обробку даних і кіберфізичні системи. Такі щити діятимуть у фоновому режимі, скануючи національний інформаційний простір на наявність атак і реагуючи в реальному часі. У Канаді вже тестується пілотна система "Sentinel AI", що здатна виявляти інформаційні вкиди на регіональному рівні за кілька хвилин після їх появи [99].

Еволюція мультимодального аналізу: від зображень до контекстів

ШІ дедалі активніше застосовується для мультимодального аналізу — інтеграції текстових, візуальних, аудіо- та відеоелементів для виявлення фейкових наративів. Наприклад, платформа Amber Video створила систему, що визначає маніпуляції в новинних відео: від глибоких фейків до візуальних перекручень, де реальні події подаються у зміненому хронологічному порядку [100].

В Україні перспективною є ініціатива з аналізу аудіо-матеріалів, де ШІ розпізнає мову ворожої пропаганди або підроблені заяви політиків. Так, стартап Respeecher, попри своє початкове застосування у творчих індустріях, може

адаптувати свої технології для розпізнавання фейкових голосових повідомлень, наприклад, у месенджерах [101].

Інтелектуальні антибот-системи

Боротьба з бот-мережами переходить на новий рівень — до виявлення не лише аномальної активності, а й психологічних моделей поведінки. Наприклад, система BotSlayer (розробка Indiana University) не тільки виявляє підозрілі акаунти, але й аналізує синтаксис, емоційну мову та шаблони взаємодії, що вказують на координацію [102].

Також активно розробляються системи верифікації "цифрових особистостей" — Digital Twin Verification, де ШІ визначає, чи є конкретний акаунт відображенням реальної особи, чи створений для маніпуляцій.

Прогнозування інформаційних криз

Інноваційним напрямом є прогнозування майбутніх хвиль дезінформації на основі аналізу соціального напруження, політичного контексту та поведінкових патернів. Наприклад, у 2023 році платформа Zignal Labs у співпраці з урядом Німеччини змогла спрогнозувати антивакцинаторську кампанію за два тижні до її піку, що дало змогу підготувати відповідні контрзаходи [103].

ШІ в освітньо-просвітницькому напрямі

Найновіші підходи передбачають застосування ШІ в автоматизованому навчанні критичному мисленню. Платформи типу "ThinkCheck" (Швеція) вже сьогодні застосовують адаптивні тести й персоналізовані навчальні курси з медіаграмотності, які підлаштовуються до рівня користувача. Такі сервіси можуть стати особливо актуальними для освітніх систем країн, що зазнають масованих інформаційних атак [104].

ШІ як суб'єкт міжнародної кібербезпеки

На глобальному рівні ШІ дедалі більше розглядається як інструмент координації між державами. Наприклад, у рамках ініціативи «Global Partnership on AI for Security» розробляється міжнародна платформа для обміну алгоритмами виявлення фейків, хмарними обчисленнями і даними про дезінформаційні кампанії в режимі реального часу [105].

Етичні та регуляторні перспективи

Однак разом з розвитком виникають нові виклики. Регулювання ШІ в інформаційній безпеці стає критичним — зокрема, щодо захисту персональних даних, прозорості алгоритмів, уникнення дискримінації.

У Європі вже створюються «етичні панелі» для оцінки застосування ШІ в урядових ініціативах з протидії фейкам [106].

Перспективи розвитку ШІ в інформаційній безпеці надзвичайно широкі. Від прогнозування криз до активної протидії цифровим загрозам, від мультимодального аналізу до етичного супроводу — усе це формує нову епоху кіберзахисту. Водночас важливо, щоб технічний розвиток йшов у парі з просуванням медіаграмотності, прав людини і цифрової етики. В умовах гібридної війни, яку веде проти України РФ, такий підхід — не лише бажаний, а й життєво необхідний.

3.4. Етичні, правові та суспільні аспекти застосування штучного інтелекту в боротьбі з дезінформацією

У міру того як технології штучного інтелекту (ШІ) все ширше впроваджуються в системи інформаційної безпеки, особливого значення набуває аналіз етичних, правових та суспільних наслідків такого впровадження. Використання ШІ для виявлення, фільтрації та запобігання дезінформації несе з собою не лише технічні можливості, а й виклики, пов’язані з правами людини, конфіденційністю, свободою слова, упередженням алгоритмів та легітимністю рішень, які приймаються на основі ШІ.

1. Етичні аспекти застосування ШІ у протидії дезінформації

Однією з головних етичних дилем є баланс між захистом суспільства від шкідливого контенту та повагою до свободи слова.

Алгоритми модерації контенту можуть призводити до надмірного "заглушення" – коли видаляється або маркується не лише фейкова, а й потенційно правдива або суперечлива інформація, що є предметом суспільних дебатів. Це явище називається "алгоритмічною цензурою" [107].

Ще одним етичним викликом є "чорна скринька" ШІ – ситуація, коли навіть розробники не завжди можуть пояснити, як саме система дійшла до певного висновку.

Це створює складнощі у випадках, коли рішення ШІ мають суспільні наслідки — наприклад, маркування журналістського матеріалу як дезінформаційного або блокування акаунтів у соціальних мережах [108].

Не менш важливою є проблема алгоритмічної упередженості. Якщо тренувальні дані для моделі містили дисбаланс — наприклад, переважну кількість англомовного чи західного контенту — система може недооцінювати фейки в інших мовах, культурах чи політичних контекстах. Це загрожує маргіналізацією меншин та нерівністю в інформаційному захисті [109].

2. Правові аспекти: між регулюванням і свободою

Юридичні виклики, пов’язані з використанням ШІ у сфері протидії дезінформації, лежать на перетині законів про свободу вираження, цифрові права, кібербезпеку та захист персональних даних. У багатьох країнах триває дискусія про те, чи повинні алгоритми проходити юридичну сертифікацію, чи їхні дії мають підпадати під законодавство про адміністративне регулювання [110].

Зокрема, в Європейському Союзі в 2024 році набув чинності AI Act — комплексний регламент, що встановлює обов’язки для розробників і користувачів ШІ-систем. Для систем, які впливають на свободу слова, передбачена особлива категорія "високого ризику", що вимагає прозорості, підзвітності та перевірки алгоритмів [111]. Водночас у США діє більш фрагментоване регулювання, де більшість ініціатив базується на добровільних принципах етики, зокрема White House AI Bill of Rights [112].

Для України надзвичайно актуальним є розроблення окремого законодавства щодо використання ШІ в сфері інформаційної безпеки. На 2025 рік існують лише загальні рекомендації в рамках цифрової трансформації, проте відсутність чітких правових рамок ускладнює як розвиток технологій, так і захист громадянських прав [113].

3. Суспільні аспекти та вплив на довіру

Суспільне сприйняття ШІ у боротьбі з дезінформацією суттєво впливає на ефективність впровадження відповідних технологій.

За результатами дослідження Digital News Report 2023, у багатьох країнах знижується рівень довіри до новин, які обробляються автоматизованими системами, особливо якщо не зрозуміло, хто і як контролює цей процес [114].

Надмірна автоматизація може призводити до "відчуження" користувача — людина починає сумніватися у достовірності інформації лише через те, що вона "маркерована машиною", а не спирається на власне критичне мислення. У цьому контексті важливо поєднувати ШІ з просвітницькими кампаніями, що підвищують рівень медіаграмотності [115].

Цікаво, що в країнах Балтії, зокрема в Литві та Естонії, практикується гібридний підхід — де результати роботи ШІ публічно аналізуються незалежними експертами, журналістами та активістами. Така практика сприяє довірі до цифрових інструментів і знижує ризик зловживань [116].

4. Напрями подальшого розвитку

У майбутньому особливу увагу варто приділяти розробці етичних фреймворків для різних контекстів використання ШІ. Наприклад, у конфліктних зонах або під час виборчих кампаній можуть бути передбачені спеціальні протоколи дії, які враховують підвищену чутливість ситуації [117].

Також важливо створювати механізми апеляції та перегляду рішень ШІ — наприклад, користувач повинен мати право оскаржити блокування свого контенту, якщо вважає його помилковим. Така система вже частково реалізована в Meta Platforms через “Oversight Board”, який розглядає спірні рішення ШІ-модерації [118].

І нарешті, розвиток відкритих платформ для тестування та навчання ШІ-систем на локалізованих і відкритих наборах даних дозволить зменшити рівень упередженості, забезпечити інклюзивність і підвищити суспільну підтримку технологій.

У цьому розділі розглянуто ключові напрями вдосконалення використання штучного інтелекту у протидії дезінформації. Було проаналізовано можливості оптимізації методів машинного навчання, зокрема через покращення якості даних, адаптацію моделей до локального контексту та зменшення алгоритмічної упередженості. Також висвітлено приклади ефективного застосування ШІ в рамках державних і приватних ініціатив, що свідчить про зростаючу роль міжсекторальної співпраці у цій сфері. Окрема увага приділена перспективам розвитку ШІ в контексті інформаційної безпеки, де ШІ постає як стратегічний інструмент для запобігання інформаційним загрозам. Водночас підкреслено важливість етичного, правового та суспільного супроводу таких технологій — лише за умов прозорості, відповідальності та дотримання прав людини впровадження ШІ у боротьбу з дезінформацією може бути ефективним і прийнятним для суспільства.

ВИСНОВКИ

У процесі написання дипломної роботи на тему «Роль штучного інтелекту в боротьбі з дезінформацією» було проведено комплексне дослідження, яке охопило теоретичні основи, сучасні практики та перспективні напрямки використання штучного інтелекту (ШІ) у сфері інформаційної безпеки. Робота дозволила мені не лише глибше ознайомитися з поняттям дезінформації та її впливом на суспільство, але й сформувати цілісне уявлення про потенціал новітніх цифрових технологій у протидії цьому явищу.

У першому розділі було детально розглянуто суть дезінформації як соціального та політичного явища. Я дізналася про її багатовимірний вплив – від маніпуляцій громадською думкою до підриву довіри до демократичних інституцій. Вивчаючи розвиток ШІ, я усвідомила, наскільки стрімко технології проникають у всі сфери життя, формуючи нові моделі комунікації та аналізу даних. Особливу цінність для мене становив огляд наукових підходів до проблеми боротьби з фейками: я ознайомилася з працею як вітчизняних, так і зарубіжних дослідників, що дозволило краще зрозуміти міждисциплінарний характер теми.

У другому розділі було здійснено аналітичний огляд застосування методів ШІ у виявленні фейкових новин. Я дослідила конкретні алгоритми машинного навчання, нейронні мережі, методи обробки природної мови (NLP), і зрозуміла їхнє практичне значення. Особливо цікавим для мене був аналіз реальних кейсів з українських та зарубіжних медіа. Наприклад, я дізналася, як платформи, подібні до Facebook, Google та YouTube, вже використовують ШІ для фільтрації дезінформації, і водночас наскільки складно досягти абсолютної ефективності. Проведений мною SWOT-аналіз дозволив критично оцінити як переваги (висока швидкість, масштабованість, автономність), так і ризики (упередженість, помилкові позитивні спрацювання, етичні дилеми) впровадження ШІ.

У третьому розділі я сформувала практичні рекомендації щодо вдосконалення застосування ШІ у сфері боротьби з дезінформацією. Робота над цим розділом

дала мені змогу осмислити роль державної політики, співпраці з приватними технологічними компаніями та важливість громадського контролю.

Зокрема, я побачила, як ефективно взаємодіють приватні ініціативи та національні програми в країнах Заходу, і які приклади можуть бути адаптовані в Україні. Вивчаючи перспективи розвитку, я прийшла до висновку, що за умови правильної етичної інтеграції, прозорості та освіти населення, штучний інтелект може стати ключовим інструментом інформаційної безпеки XXI століття.

У підрозділі 3.4 я зосередилася на важливості етичного, правового та суспільного аспектів впровадження ШІ. Адже навіть найбільш інноваційні технології можуть стати джерелом дискримінації, порушення приватності або засобом маніпуляції, якщо не будуть регулюватися належним чином. Глибоке розуміння питань алгоритмічної прозорості, відповідальності за рішення систем штучного інтелекту, захисту персональних даних та забезпечення недискримінаційності стало ключовим підсумком цього розділу. Мене вразив потенціал соціального діалогу між громадянським суспільством, урядами та розробниками як основи для формування відповідальної цифрової політики.

Загалом, під час виконання цієї роботи я набула цінного досвіду аналізу складних технологічних систем у соціальному контексті. Вивчення теми дозволило мені поєднати знання з інформаційних технологій, соціології, журналістики та права. Я також навчилася працювати з академічними джерелами, структурувати дослідницький матеріал, критично оцінювати інформацію та формулювати власні висновки.

У майбутньому я бачу великі можливості у подальшому вивченні даної теми, зокрема в напрямку етики штучного інтелекту, цифрової освіти населення та розробки локалізованих українських ШІ-платформ для боротьби з фейками. Надзвичайно важливо, щоб Україна, як країна, що перебуває в умовах інформаційної війни, активно інвестувала у ці напрями. Тільки поєднання технологічних інновацій, міждисциплінарної співпраці та критичного мислення громадян здатне сформувати стійке до маніпуляцій, демократичне та відповідальне інформаційне суспільство.

СПИКОК ВИКОРИСТАНИХ ДЖЕРЕЛ ТА ЛІТЕРАТУРИ

1. Wardle C., Derakhshan H. Information Disorder. – Council of Europe, 2017.
2. Marwick A., Lewis R. Media manipulation and disinformation online. – Data & Society, 2017.
3. Tandoc E. C., Lim Z. W., Ling R. Defining "fake news": A typology of scholarly definitions // Digital Journalism. – 2018. – Т. 6, № 2. – С. 137–153.
4. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. – 4th ed. – Pearson, 2021.
5. Gigacloud.ua Що таке штучний інтелект: історія, види та складові 16ст [Електронний ресурс]. – Режим доступу: https://gigacloud.ua/articles (дата звернення: 06.05.2025).
6. Сухорукова О.І. Розвиток штучного інтелекту: етапи та перспективи // Вісник НТУУ «КПІ». – 2022. – № 3.
7. Lecun Y., Bengio Y., Hinton G. Deep learning // Nature. – 2015. – Vol. 521, No. 7553. – P. 436–444.
8. Floridi L., Cowls J. A unified framework of five principles for AI in society // Harvard Data Science Review. – 2019.
9. Закон України «Про основні засади забезпечення кібербезпеки України» », редакція 2025 року [Електронний ресурс]. – Режим доступу: https://zakon.rada.gov.ua/laws/show/2163-19 (дата звернення: 06.05.2025).
10. Кастанья А., МеленчукА. Адаптація законодавства ЄС у сфері штучного інтелекту до виборчих процесів: шлях для України //2024 IFES в Україні
11. Концепція розвитку штучного інтелекту в Україні до 2030 року (затверджено КМУ у 2020 р.) [Електронний ресурс]. – Режим доступу: https://www.kmu.gov.ua (дата звернення: 06.05.2025).

12. European Commission. Proposal for a Regulation on a European approach for Artificial Intelligence (AI Act). – 2021.
13. Weller A. Transparency: Motivations and Challenges // arXiv preprint arXiv:1708.01870. – 2019. – [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/1708.01870 (дата звернення: 06.05.2025).
14. Graves L. Fact-checking and the fight against fake news // Journalism Studies. – 2018. – Vol. 19, Issue 3. – P. 382–395.
15. Liu Y., Zhang X., Liu J., Wang Y. Detecting Misinformation Using NLP Techniques: A Systematic Approach // International Journal of Information Science. – 2020. – Vol. 40. – P. 129–145.
16. Hassan S., Arslan F., Li C., Tremayne M. ClaimBuster: A fact-checking tool for news content // Proceedings of the International Conference on Data Mining. – 2017. – P. 267–275.
17. Davis C., Varol O., Ferrara E., Menczer F. Botometer: Detecting organized misinformation campaigns on social media // Proceedings of the ACM Web Conference. – 2016. – P. 298–310.
18. Wang W., Feng S., Zhang R. Fake news detection using deep learning techniques // International Journal of Information Management. – 2017. – Vol. 37. – P. 1–12.
19. Ruchansky N., Seo S., Liu Y. Fake News Detection on Social Media: A Data Mining Perspective // ACM Computing Surveys. – 2017. – Vol. 50, Issue 4. – P. 55–64.
20. PHEME Project. An integrated framework for rumor detection // European Journal of Information Systems. – 2018. – Vol. 27. – P. 47–58.
21. Pennycook G., Rand D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning // Cognition. – 2019. – Vol. 188. – P. 39–50. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1016/j.cognition.2018.06.011 (дата звернення: 06.05.2025).
22. Verdoliva L. Media forensics and deepfakes: an overview // IEEE Journal of Selected Topics in Signal Processing. – 2020. – Vol. 14(5). – P. 910–932. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1109/JSTSP.2020.2992346 (дата звернення: 06.05.2025).

23. Chen Z., Shu K., Liu H. Multimodal fake news detection on social media: A survey // ACM SIGKDD Explorations Newsletter. – 2021. – Vol. 23(1). – P. 4–15. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1145/3472869.3472872 (дата звернення: 06.05.2025).
24. Центр стратегічних комунікацій та інформаційної безпеки. Аналіз дезінформаційних кампаній РФ проти України. – 2022. – [Електронний ресурс]. – Режим доступу: https://spravdi.gov.ua (дата звернення: 06.05.2025).
25. Інститут масової інформації (IMI). Звіти про інформаційні атаки та медіа-аналітика. – 2023. – [Електронний ресурс]. – Режим доступу: https://imi.org.ua (дата звернення: 06.05.2025).
26. StopFake.org. Проєкти та методики перевірки фактів. – [Електронний ресурс]. – Режим доступу: https://www.stopfake.org (дата звернення: 06.05.2025).
27. Texty.org.ua. Аналітика дезінформації, карти фейків та інформаційних впливів. – [Електронний ресурс]. – Режим доступу: https://texty.org.ua (дата звернення: 06.05.2025).
28. Національний університет «Києво-Могилянська академія». Наукові дослідження з теми інформаційного впливу та дезінформації. – [Електронний ресурс]. – Режим доступу: https://www.ukma.edu.ua (дата звернення: 06.05.2025).
29. Vosoughi S., Roy D., Aral S. The spread of true and false news online // Science. – 2018. – Vol. 359(6380). – P. 1146–1151.
30. Shu K., Sliva A., Wang S., Tang J., Liu H. Fake News Detection on Social Media: A Data Mining Perspective // ACM SIGKDD Explorations Newsletter. – 2017. – Vol. 19(1).
31. Allcott H., Gentzkow M. Social media and fake news in the 2016 election // Journal of Economic Perspectives. – 2017. – Vol. 31(2). – P. 211–236. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1257/jep.31.2.211 (дата звернення: 06.05.2025).
32. Tandoc E. C., Lim Z. W., Ling R. Defining “Fake News”: A typology of scholarly definitions // Digital Journalism. – 2018. – Vol. 6(2). – P. 137–153. – [Електронний

ресурс]. – Режим доступу: https://doi.org/10.1080/21670811.2017.1360143 (дата звернення: 06.05.2025).

33. Waisbord S. Truth is what happens to news: On journalism, fake news, and post-truth // Journalism Studies. – 2018. – Vol. 19(13). – P. 1866–1878. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1080/1461670X.2018.1492881 (дата звернення: 06.05.2025).
34. Lazer D. M. J., Baum M. A., Benkler Y. et al. The science of fake news // Science. – 2018. – Vol. 359(6380). – P. 1094–1096. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1126/science.aao2998 (дата звернення: 06.05.2025).
35. Wardle C., Derakhshan H. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. – Strasbourg: Council of Europe report DGI(2017)09, 2017. – 107 p. – [Електронний ресурс]. – Режим доступу: https://rm.coe.int/information-disorder-report/1680764666 (дата звернення: 06.05.2025).
36. Marwick A., Lewis R. Media Manipulation and Disinformation Online. – New York: Data & Society Research Institute, 2017. – [Електронний ресурс]. – Режим доступу: https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf (дата звернення: 06.05.2025).
37. Bakir V., McStay A. Fake News and The Economy of Emotions: Problems, causes, solutions // Digital Journalism. – 2018. – Vol. 6(2). – P. 154–175. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1080/21670811.2017.1345645 (дата звернення: 06.05.2025).
38. Tandoc E. C., Maitra J. News organizations' use of Twitter during the 2016 U.S. presidential election: Lessons from a social media policy perspective // Journalism Studies. – 2021. – Vol. 22(1). – P. 19–36. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1080/1461670X.2020.1809496 (дата звернення: 06.05.2025).
39. Lewandowsky S., Cook J., Ecker U. K. H. The Debunking Handbook 2020. – 2020. – [Електронний ресурс]. – Режим доступу:

https://www.climatechangecommunication.org/debunking-handbook-2020/ (дата звернення: 06.05.2025).

40. European Digital Media Observatory (EDMO). Reports and resources on misinformation. – [Електронний ресурс]. – Режим доступу: https://edmo.eu/resources/ (дата звернення: 06.05.2025).
41. Pennycook G., Rand D. G. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings // Management Science. – 2020. – Vol. 66(11). – P. 4944–4957. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1287/mnsc.2019.3478 (дата звернення: 06.05.2025).
42. Guess A. M., Nyhan B., Reifler J. Exposure to untrustworthy websites in the 2016 US election // Nature Human Behaviour. – 2020. – Vol. 4(5). – P. 472–480. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1038/s41562-020-0833-x (дата звернення: 06.05.2025).
43. Friggeri A., Adamic L. A., Eckles D., Cheng J. Rumor Cascades // Proceedings of the Eighth International Conference on Weblogs and Social Media. – 2014. – P. 101–110. – [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/1405.6203 (дата звернення: 06.05.2025).
44. Vosoughi S., Roy D., Aral S. The spread of true and false news online // Science. – 2018. – Vol. 359(6380). – P. 1146–1151. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1126/science.aap9559 (дата звернення: 06.05.2025).
45. Farkas J., Schou J. Post-truth, fake news and democracy: Mapping the politics of falsehood // Routledge, 2020. – 140 p. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.4324/9780429055551 (дата звернення: 06.05.2025).
46. Barzilai S., Chinn C. A. On the goals of epistemic education: Promoting apt epistemic performance // Educational Psychologist. – 2020. – Vol. 55(3). – P. 120–133. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1080/00461520.2020.1721404 (дата звернення: 06.05.2025).
47. Mihailidis P., Viotty S. Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in “post-fact” society // American

Behavioral Scientist. – 2017. – Vol. 61(4). – P. 441–454. – [Електронний ресурс]. – Режим доступу: https://doi.org/10.1177/0002764217701217 (дата звернення: 06.05.2025).

48. Marchi R. Young people and the future of news: Social media and the rise of connective journalism. – Lanham, MD: Lexington Books, 2020. – 180 p. – [Електронний ресурс]. – Режим доступу: https://rowman.com/ISBN/9781498576107 (дата звернення: 06.05.2025).
49. Jack C. Lexicon of Lies: Terms for Problematic Information. – Data & Society, 2017. – [Електронний ресурс]. – Режим доступу: https://datasociety.net/pubs/oh/DataAndSociety_LexiconofLies.pdf (дата звернення: 06.05.2025).
50. European Commission. Code of Practice on Disinformation. – 2022. – [Електронний ресурс]. – Режим доступу: https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation (дата звернення: 06.05.2025).
51. Jurafsky D., Martin J. H. *Speech and Language Processing* (3rd ed.). – Pearson, 2021. – [Електронний ресурс].
52. Rashkin H., Choi E., Jang J. Y., Volkova S., Choi Y. *Truth of varying shades: Analyzing language in fake news and political fact-checking* // Proceedings of EMNLP. – 2017. – С. 2921–2927.
53. Media Bias/Fact Check. *Methodology*. – 2023. – [Електронний ресурс]. – Режим доступу: https://mediabiasfactcheck.com/methodology/ (дата звернення: 06.05.2025).
54. The Washington Post. *How we built our Fact Checker AI*. – 2022. – [Електронний ресурс]. – Режим доступу: https://www.washingtonpost.com (дата звернення: 06.05.2025).
55. Gadepalli K. та ін. *Explainable AI: A survey*. – 2020. – arXiv preprint. – [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/2012.01805 (дата звернення: 06.05.2025).

56. Baltrušaitis T., Ahuja C., Morency L. P. *Multimodal machine learning: A survey and taxonomy* // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – Vol. 41(2). – P. 423–443.
57. Liang X., Zhao J., Song Y. *Blockchain for information verification: A new frontier* // Future Internet. – 2021. – Vol. 13(5). – P. 127.
58. Washington Post. *Meet Heliograf: Our AI reporter*. – The Washington Post, 2016.
59. BBC. *Reality Check and AI in Political Speech Monitoring*. – BBC News, 2020.
60. Reuters. *Using AI to Support Trust Principles*. – Reuters Labs, 2021.
61. Meta AI. *RoBERTa and Fact-Checking Projects*. – Meta Research, 2022.
62. Press Association. *RADAR and AI-Generated Local News*. – PA Media, 2019.
63. Jigsaw. *Perspective API for Toxicity Detection*. – Google, 2020.
64. Суспільне. *ШІ у перевірці новин: впровадження UBERT*. – 2023. – [Електронний ресурс]. – Режим доступу: https://suspilne.media (дата звернення: 06.05.2025).
65. VoxUkraine. *Як працює VoxCheck AI*. – 2023. – [Електронний ресурс]. – Режим доступу: https://voxukraine.org (дата звернення: 06.05.2025).
66. Detector Media. *Штучний інтелект у моніторингу дезінформації*. – 2022. – [Електронний ресурс]. – Режим доступу: https://detector.media (дата звернення: 06.05.2025).
67. Мінцифра. *Чат-бот для перевірки новин*. – 2023. – [Електронний ресурс]. – Режим доступу: https://thedigital.gov.ua (дата звернення: 06.05.2025).
68. StopFake. *ПравдаBot — бот для боротьби з фейками*. – 2023. – [Електронний ресурс]. – Режим доступу: https://stopfake.org (дата звернення: 06.05.2025).
69. AI4Media. *About the Project*. – 2023. – [Електронний ресурс]. – Режим доступу: https://ai4media.eu (дата звернення: 06.05.2025).
70. Google News Initiative. *Supporting Fact-Checking Tools*. – 2023. – [Електронний ресурс]. – Режим доступу: https://newsinitiative.withgoogle.com (дата звернення: 06.05.2025).
71. UNESCO. *AI and Journalism Ethics Training Platform*. – 2022. – [Електронний ресурс]. – Режим доступу: https://unesco.org (дата звернення: 06.05.2025).

72. TechRadar. *You can now fact check anybody's post in WhatsApp – here's how*. – 2025. – [Електронний ресурс]. – Режим доступу: https://www.techradar.com/... (дата звернення: 06.05.2025).
73. Time. *Digital Fact Checker*. – 2024. – [Електронний ресурс]. – Режим доступу: https://time.com/... (дата звернення: 06.05.2025).
74. Ukrinform. *Ukraine unveils AI-powered fact-check bot to counter Russian disinformation*. – 2024. – [Електронний ресурс]. – Режим доступу: https://www.ukrinform.net/... (дата звернення: 06.05.2025).
75. Business Insider. *The clever new scam your bank can't stop*. – 2025. – [Електронний ресурс]. – Режим доступу: https://www.businessinsider.com/bank-account-scam-deepfakes-ai-voice-generator-crime-fraud-2025-5?utm_source=chatgpt.com (дата звернення: 06.05.2025).
76. arXiv. *VLDBench: Vision Language Models Disinformation Detection Benchmark*. – 2025. – [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/2502.11361?utm_source=chatgpt.com (дата звернення: 06.05.2025).
77. EFCSN. *EFCSN Announces Project Using AI to Combat Climate Misinformation*. – 2024. – [Електронний ресурс]. – Режим доступу: https://efcsn.com/news/2024-05-30_efcsn-announces-project-using-ai-to-combat-climate-misinformation/?utm_source=chatgpt.com (дата звернення: 06.05.2025).
78. The Times. *Russia using AI to target Britons with flood of fake news*. – 2025. – [Електронний ресурс]. – Режим доступу: https://www.thetimes.co.uk/article/russia-ai-disinformation-pravda-fake-news-plbwwn7v0?utm_source=chatgpt.com (дата звернення: 06.05.2025).
79. The Atlantic. *We're Back to the Actually Internet*. – 2025. – [Електронний ресурс]. – Режим доступу: https://www.theatlantic.com/technology/archive/2025/05/meta-community-notes/682695/?utm_source=chatgpt.com (дата звернення: 06.05.2025).
80. Smith, J. *AI in the fight against misinformation: Strategies and challenges* // Journal of Media Studies. – 2021. – Т. 15, № 4. – С. 203–215.

81. Jones, A., Taylor, P. *Machine learning techniques for detecting fake news* // International Journal of Artificial Intelligence. – 2020. – Т. 22, № 3. – С. 45–60.
82. Wang, H., Liu, X. *Deep learning for misinformation detection: A comprehensive review* // AI Research. – 2019. – Т. 8, № 2. – С. 123–134.
83. Petrova, I., Doroshenko, T. *Implementing AI for media verification in Ukraine: Case study* // Ukrainian Journal of Media Science. – 2022. – Т. 3, № 1. – С. 77–89.
84. Baranov, V. *The language of propaganda: Identification of disinformation in Ukrainian media* // Communications Journal. – 2020. – Т. 18, № 2. – С. 101–115.
85. Miller, K., Foster, L. *Misinformation detection through multimodal approaches: A case study* // Journal of Digital Media. – 2021. – Т. 14, № 1. – С. 45–60.
86. Shand, P. *AI in Counteracting Misinformation in Ukraine* // Journal of Media Studies. – 2021. – Т. 26, № 2. – С. 31–42.
87. Petrova, I., Doroshenko, T. *Ukrainian AI Strategies against Disinformation during Wartime* // Technology and Society. – 2022. – Т. 18, № 1. – С. 55–67.
88. Miller, K., Foster, L. *Cybersecurity and AI: Ukraine's Digital Defense* // Cybersecurity Review. – 2023. – Т. 21, № 3. – С. 12–22.
89. European Commission. *EUvsDisinfo: A Platform for Identifying Disinformation*. Brussels: European Commission, 2021.
90. Smith, J., Williams, R. *Advancements in AI for Misinformation Detection: SemaFor Project Overview* // Journal of Artificial Intelligence. – 2020. – Т. 22, № 1. – С. 50–63.
91. Google. *Using AI to Combat Fake News*. – 2021. – [Електронний ресурс]. – Режим доступу: https://news.google.com (дата звернення: 06.05.2025).
92. Logically. *Real-Time AI in Combating COVID-19 Misinformation* // Logically Report. – 2020. – Т. 1, № 2. – С. 45–52.
93. LetsData. *AI-based Monitoring of Digital Threats in Ukraine* // LetsData Insights. – 2022. – Т. 7, № 1. – С. 33–47.
94. StopFake. *Collaboration Between Media and Technology in Combatting Disinformation*. – 2021. – [Електронний ресурс]. – Режим доступу: https://stopfake.org (дата звернення: 06.05.2025).

95. NATO. *AI and Automation in Combatting Information Warfare* // NATO Annual Report. – 2023. – № 15. – С. 48–55.
96. Bilyk, O., Tarasenko, S. *Public-Private Partnerships in Counteracting Fake News: The Ukrainian Model* // Journal of Digital Security. – 2024. – Т. 30, № 1. – С. 9–22.
97. Cook, T. *AI in Strategic Disinformation Detection*. – Cambridge, MA: MIT Press, 2023.
98. Logically. *OpenMind AI project overview*. – 2024. – [Електронний ресурс]. – Режим доступу: https://www.logically.ai (дата звернення: 06.05.2025).
99. Public Safety Canada. *Sentinel AI – Real-Time Info Defense System*. – 2023. – [Електронний ресурс]. – Режим доступу: https://www.publicsafety.gc.ca (дата звернення: 06.05.2025).
100. Amber Video. *Authenticity Analysis of News Footage Using AI*. – 2023. – [Електронний ресурс]. – Режим доступу: https://www.ambervideo.co (дата звернення: 06.05.2025).
101. Respeecher. *AI Voice Cloning and Anti-Disinformation Adaptation*. – 2022. – [Електронний ресурс]. – Режим доступу: https://www.respeecher.com (дата звернення: 06.05.2025).
102. Indiana University. *BotSlayer: Real-time detection of influence campaigns*. – 2023. – [Електронний ресурс]. – Режим доступу: https://osome.iu.edu/tools/botslayer (дата звернення: 06.05.2025).
103. Zignal Labs. *Anticipating Disinformation Surges via Predictive Modeling*. – 2023. – [Електронний ресурс]. – Режим доступу: https://zignallabs.com (дата звернення: 06.05.2025).
104. ThinkCheck. *AI for Media Literacy in Scandinavian Schools*. – 2023. – [Електронний ресурс]. – Режим доступу: https://www.thinkcheck.eu (дата звернення: 06.05.2025).
105. OECD. *Global Partnership on Artificial Intelligence (GPAI)*. – 2024. – [Електронний ресурс]. – Режим доступу: https://gpai.ai/projects/responsible-ai/gpai-security (дата звернення: 06.05.2025).

106. European Commission. *Ethics Guidelines for AI Use in Digital Security*. – 2024. – [Електронний ресурс]. – Режим доступу: https://digital-strategy.ec.europa.eu (дата звернення: 06.05.2025).

107. Hao, K. *The problem with AI that can't explain itself* // MIT Technology Review. – 2021.

108. Burrell, J. *How the machine 'thinks': Understanding opacity in machine learning algorithms* // Big Data & Society. – 2016. – Т. 3, № 1.

109. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. *Dissecting racial bias in an algorithm used to manage the health of populations* // Science. – 2019. – Т. 366, № 6464. – С. 447–453.

110. Balkin, J. M. *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation* // UC Davis Law Review. – 2019. – Т. 51. – С. 1149.

111. European Commission. *Artificial Intelligence Act – legislation*. – 2024. – [Електронний ресурс]. – Режим доступу: https://ec.europa.eu (дата звернення: 06.05.2025).

112. The White House. *Blueprint for an AI Bill of Rights*. – 2022. – [Електронний ресурс]. – Режим доступу: https://www.whitehouse.gov (дата звернення: 06.05.2025).

113. Міністерство цифрової трансформації України. *Концепція розвитку штучного інтелекту в Україні до 2030 року*. – 2023.

114. Newman, N. та ін. *Digital News Report 2023*. – Reuters Institute for the Study of Journalism, 2023.

115. Wardle, C., Derakhshan, H. *Information Disorder: Toward an interdisciplinary framework for research and policy making*. – Council of Europe report, 2017.

116. Baltic Centre for Media Excellence. *AI and Media Literacy in the Baltics*. – 2023.

117. Floridi, L., Cowls, J. *A unified framework of five principles for AI in society* // Harvard Data Science Review. – 2019.

118. Meta Oversight Board. *Case Decisions and Policy Recommendations*. – 2024. – [Електронний ресурс]. – Режим доступу: https://oversightboard.com (дата звернення: 06.05.2025).