Міністерство освіти і науки України
Харківський національний університет імені В. Н. Каразіна

Навчально-науковий інститут соціології та медіакомунікацій

Кафедра медіакомунікацій

До захисту допущено
кафедрою медіакомунікацій, протокол № ____ від _____

В. о. завідувача кафедри _____ Лілія ЗМІЙ
(підпис) (ім'я, прізвище)

«___»_____ 2025 р.

Кваліфікаційна робота
(інноваційний магістерський проєкт)
здобувача другого (магістерського) рівня вищої освіти

ШТУЧНИЙ ІНТЕЛЕКТ ЯК ІНСТРУМЕНТ СТВОРЕННЯ, ПОШИРЕННЯ ТА ВИЯВЛЕННЯ DEEPFAKES

Спеціальність 061 Журналістика
(код та найменування спеціальності)

Освітня програма Аудіовізуальні медіа та цифрова журналістика
(назва освітньої програми)

Виконавець _____ Володимир ГАРМАШ
(підпис) (ім'я, прізвище)

Науковий керівник _____ Лідія СТАРОДУБЦЕВА
(підпис) (ім'я, прізвище)

Харків – 2025

ЗМІСТ

ВСТУП

Обґрунтування вибору теми дослідження. За останній час технології штучного інтелекту ще більш актуалізувалися, ніж це було під час написання моєї попередньої роботи з цієї теми рік тому. Штучний інтелект, своєю чергою, став широко інтегруватися у звичайне людське життя: чат боти, голосові помічники, асистенти, цілі системи, що займаються генерацією зображень, відео- та фото контенту. Між державами, навіть правильніше сказати, між наддержавними утвореннями і спілками конкретно ставиться мета певного технологічного суверенітету в цій галузі. У США — ChatGPT і Gemini, у Європі — Mistral, у Китаї — Deepseek. Ці досягнення ведуть не лише до значного технологічного прогресу, а й до появи нових викликів, пов'язаних з етикою, безпекою та довірою до цифрового контенту.

Одним із найбільш обговорюваних і суперечливих продуктів ШІ став deepfake — синтетичний медіаконтент, що створюється за допомогою нейромережевих алгоритмів. Його можливості викликають водночас захоплення та занепокоєння: з одного боку — це потужний інструмент у сфері розваг, кіновиробництва та освіти, а з іншого — пряма загроза дезінформації, політичним маніпуляціям, вторгненням у приватне життя, та навіть правам людини.

Ступінь вивченості теми. Deepfake — це не просто технологічна новинка, а складний феномен, який потребує міждисциплінарного аналізу. У контексті нашого дослідження особливе значення мають роботи , що були присвячені даній тематиці, вітчизняних дослідників і науковців. Хоча рівень вивченості deepfakes в Україні залишає бажати кращого, необхідно згадати дослідників, на висновках яких будувалася наша робота. Як говорив сер Ісаак Ньютон, «якщо я і бачив більше за інших, то тільки лише тому, що стояв на плечах гігантів». Перш за все, необхідно згадати основоположну працю Ксенії Юртаєвої, робота якої була опублікована у Віснику криміналістичної асоціації України, — саме на цій праці, а також на серії інших праць і дисертаційних досліджень з кримінології, присвячених правовим ризиком поширення deepfakes, побудована юридична

компонента даного дослідження [5] З прикладної точки зору, вважаємо за необхідне також виділити роботу Максима Прищепи «Deepfake як новий небезпечний вид кіберзброї», що була опублікована Київським політехнічним інститутом імені Ігоря Сікорського [3]. Ця робота стала однією з небагатьох україномовних, на яку ми звернули особливу увагу, досліджуючи політичні наслідки та використання deepfakes під час виборчого процесу та знакових подій, таких, як, наприклад, «ситуативні» deepfakes, що були створені з метою дискредитації українських політичних діячів під час російсько-української війни.

У сфері аналітики фінансових ризиків необхідно згадати дослідження Марини Рябоконь, викладачки Київського інституту бізнесу та технологій, за темою «Вплив технології Deepfake на Fintech сектор: поточний стан в Україні та стратегії запобігання кіберзлочинності» [4] . У сектор фінансів, у якому й так існує величезна невизначеність у нинішньому часі, ще й додається загроза deepfakes. Цей аспект, щоправда з іншої частини земної кулі, висвітлюється в нашій праці в наступних розділах, а саме у як завдяки deepfakes можна втратити мільйони доларів і такий випадок вже стався в історії, що, без жодного сумніву, додає ваги цим працям і дослідженням, бо ці вони досліджують реалії, а не гіпотетичні загрози. Даною проблематикою займались науковці Баумейстер А. , Юртаєва К. В. , Рябокінь М. В. , Котух Є. В. , Папилев О. І. , Прищепа М. О. , Качинський А. Б.

Західний науковий світ багатший на роботи, пов'язані з deepfakes. У цьому контексті необхідно згадати, перш за все, Ніка Бострома (Nick Bostrom) з його основоположною працею «Штучний інтелект. Етапи. Загрози. Стратегії», що вийшла друком ще 2016-го року [10] . Для нашого дослідження велику значущість має також одна з небагатьох робіт, написаних про використання deepfakes з погляду медицини та можливості лікування і створення спеціальних тренажерів за допомогою deepfakes для людей, які страждають на проблеми з пам'яттю, за авторством американських дослідників Сари Хоек (Sara Hoek), Сюзанни Метцелар (Suzanne Metselaar) та інших . Ця робота послужила деякою підмогою для написання розділу нашої роботи, присвяченого соціально-етичним аспектам

явища deepfakes. Дослідники, до праць яких я звертався під час написання цієї роботи : Colaner N. , Quinn M. , Cole S. , Copeland J. ,Devlin K. ,Cheetham J.,Farid H.,Zakharov E.,Goldstine H. H.,Kaur S.,Kuklychev Y.,Lipkowitz,Noek S.,O'Brien S. R.,Perez S.,Punnappurath A., Brown M. S.

Однак попри те, що окремі аспекти, на нашу думку, особливо важливо розглядати не тільки самі технології створення deepfake, а й механізми їхнього поширення, методи виявлення, а також те, як різні держави і суспільства реагують на цю загрозу. Усе це і зумовлює необхідність проведення нашого дослідження.

Мета та завдання дослідження. *Метою* дослідження є визначення специфіки технологій штучного інтелекту, що застосовуються для створення, поширення та виявлення deepfakes, на основі аналізу їх технічних особливостей, соціального впливу та етико-правових ризиків, пов'язаних з їх використанням у масових комунікаціях.

Для того, щоб досягти поставленої мети, ми маємо вирішити такі *завдання*:

– обґрунтувати концептуальні вектори аналіз штучного інтелекту як інструменту створення, поширення та виявлення deepfakes; Розкрити сутність поняття «deepfake», простежити історію його виникнення та здійснити аналіз сучасного стану наукової розробки цієї проблематики в міждисциплінарному дискурсі;

– охарактеризувати концептуальні та алгоритмічні основи функціонування штучного інтелекту в контексті генерації синтетичного медіаконтенту, зокрема проаналізувати роботу автоенкодерів, згорткових нейронних мереж та технологій синтезу голосу;

– дослідити соціально-етичні виміри феномена deepfakes крізь призму теорії симулякрів та різних підходів до осмислення штучного інтелекту (семіотичного та біологічного), виявивши ключові етичні ризики для суспільства;

– проаналізувати стан правового регулювання створення та поширення deepfakes у міжнародному (США, ЄС) та національному законодавстві, а також оцінити опосередкований вплив цієї технології на економічну безпеку України;

– здійснити емпіричний аналіз резонансних кейсів використання deepfakes в українському інформаційному просторі та виявити тенденції їх застосування у високотехнологічній злочинності;

– реалізувати авторський експеримент зі створення deepfake-контенту для демонстрації доступності технології та, на основі отриманих даних, розробити практичні рекомендації для медіафахівців щодо ідентифікації та верифікації синтетичного контенту.

Об'єкт дослідження – процеси створення, поширення та виявлення deepfake-контенту за допомогою технологій штучного інтелекту.

Предмет дослідження – технічні засоби, алгоритми та інструменти штучного інтелекту, що використовуються для генерації та детекції deepfakes; соціальні механізми їх впливу; етичні та правові наслідки використання deepfake-технологій у сучасних комунікаційних практиках.

Методи дослідження. З огляду на міждисциплінарний характер теми, дослідження базується на інтегративному підході, що поєднує підходи інформатики, філософії, правознавства, медіазнавства, соціології та культурології. У процесі аналізу застосовано як загальнонаукові, так і спеціально-наукові методи.

До загальнонаукових методів належать аналіз, синтез, індукція, дедукція, абстрагування, моделювання та узагальнення. Вони дозволили структурувати наявні знання щодо технологій штучного інтелекту, а також виявити ключові тенденції у сфері deepfake-технологій.

Для дослідження феномена deepfakes у контексті сучасного інформаційного простору використано такі спеціально-наукові методи:

– історико-генетичний (для реконструкції етапів розвитку технологій штучного інтелекту та появи deepfakes);

– системний (для комплексного аналізу архітектури нейромережевих моделей, що застосовуються у створенні deepfakes, включаючи CNN, GAN, autoencoder-архітектури);

– порівняльно-правовий (для порівняння підходів до правового регулювання deepfake-контенту у США, країнах Європейського Союзу та Україні);

– контент-аналізу (для вивчення інформаційних повідомлень у медіа, що репрезентують використання deepfake-технологій, з метою виявлення домінантних тем, мотивацій та соціальних наративів);

– герменевтичного аналізу (для інтерпретації філософських і соціокультурних концептів, зокрема симулякрів за Ж. Бодріяром, у контексті аналізу deepfakes як феномена постправди);

– кейс-стаді (для аналізу конкретних прикладів використання deepfake технологій у політичних, культурних і комерційних контекстах).

Теоретичне підґрунтя дослідження становлять філософсько-соціологічна концепція симулякрів Жана Бодріяра (“Simulacra and Simulations”), теорія постправди Лі Макінтайра (Lee McIntyre) із Бостонського університету , концепція з семіотики візуального образу Ролана Барта (Roland Barthes) [8], теорія соціального конструювання реальності Пітера Бергера (Peter L. Berger) та Томаса Лукмана (Thomas Luckmann) [1], а також сучасні дослідження у сфері штучного інтелекту та цифрової етики (Нік Бостром , Сюзанна Хоек).

Емпіричну основу дослідження складають відкриті інтернет-джерела, наукові публікації, платформи з відкритим кодом, відеоматеріали, а також нормативно-правові документи, а саме: публікації у провідних наукових журналах та на arXiv.org, зокрема з машинного навчання, штучного інтелекту, комп’ютерного зору та обробки візуальних даних (A. Punnappurath, M. S. Brown, Few-Shot Adversarial Learning etc.); джерела у галузі медіааналітики й журналістські розслідування щодо використання deepfakes у політичних і соціальних контекстах (*BBC, Newsweek, Vice, BuzzFeed, TechCrunch*); правові аналітичні документи міжнародних організацій, що стосуються deepfake-регулювання (*Europol, NYU Journal of Legislation & Public Policy*); технічна документація та приклади застосування (*GitHub, Nvidia Maxine*).

Наукова новизна дослідження полягає визначення специфіки технологій штучного інтелекту, що застосовуються для створення, поширення та виявлення

deepfakes, на основі аналізу їх технічних особливостей, соціального впливу та етико-правових ризиків, пов'язаних з їх використанням у масових комунікаціях, а саме: проведено комплексний міждисциплінарний аналіз феномена deepfakes як технологічного, соціального-етичного та правового явища; запропоновано авторську типологізацію алгоритмів створення deepfake-контенту за функціональним принципом (заміна обличчя, синтез голосу ,модифікація міміки тощо); систематизовано сучасні концептуальні підходи до виявлення маніпуляцій візуальними та аудіо-даними; проаналізовано функціонування deepfake-технологій як симуляційної практики в дусі постструктуралістської філософії Ж. Бодріяра; виявлено основні соціально-етичні ризики використання deepfakes у медіа, політиці, кіберзлочинності та психотерапії; окреслено особливості нормативного регулювання deepfakes у США, ЄС та Україні, а також перспективи створення єдиного правового підходу до цифрової ідентичності; здійснено аналітичний огляд позитивних сценаріїв використання deepfake-технологій (освіта, медицина, інклюзія, творчість).

Практичне значення одержаних результатів дослідження полягає в можливості їхнього застосування в декількох напрямках. Матеріали і результати магістерського проекту можуть стати в пригоді для науковців (як теоретична основа для подальших досліджень у галузях штучного інтелекту, цифрової етики, кримінології, соціології медіа); розробників ШІ та медіа платформ (для створення систем виявлення deepfakes і побудови політик відповідального використання ШІ); правників і законодавців (як підґрунтя для розробки нормативно-правової бази щодо боротьби з дезінформацією та порушенням цифрових прав); представників медіа, психологів, PR-спеціалістів (у розробці ними стратегій ідентифікації та реакції на візуальні маніпуляції у цифровому просторі). Ключові положення роботи можна використовувати у освітньому процесі у межах викладання курсів з медіаграмотності, інформаційної безпеки, філософії технологій, права, ІТ та ін.

Структура та обсяг дослідження. Робота складається зі вступу, п’яти розділів (у тому числі 19-ти підрозділів), висновків, списку використаних джерел (60 найменувань, з них 47 – іноземними мовами, 6 – досліджувані джерела) та ілюстративних додатків. Загальний обсяг дослідження – 94 сторінки (з них основна частина – 72 сторінки).

РОЗДІЛ 1
ШТУЧНИЙ ІНТЕЛЕКТ ЯК ІНСТРУМЕНТ СТВОРЕННЯ, ПОШИРЕННЯ ТА ВИЯВЛЕННЯ DEEPFAKES

У межах дослідження доцільно здійснити причинно-наслідковий аналіз технологічних передумов виникнення deepfakes. Слід зазначити, що саме стрімкий прогрес у сфері нейронних мереж став каталізатором появи та поширення deepfakes як інструменту. За своєю сутністю ці технології є еволюційним продовженням гіпотез та постулатів, сформульованих ще у 50-х роках XX століття.

Як зазначає американський дослідник історії комп’ютерних технологій Герман Голдстейн, функціонування сучасних нейронних мереж, що генерують deepfakes, ґрунтується на архітектурі Джона фон Неймана — угорсько-американського фізика і математика, який у середині минулого століття розробив фундаментальну концепцію побудови обчислювальних машин [22, p. 24]. Ключовою особливістю цієї архітектури, за словами Голдстейна, є принцип зберігання команд для виконання безпосередньо в оперативній пам’яті пристрою.

Варто зауважити, що ранні електронно-обчислювальні машини не використовували подібні принципи організації пам’яті. Наприклад, під час Другої світової війни для дешифрування повідомлень німецької машини «Енігма» (Enigma – з нім. «Загадка») союзниками були залучені видатні математики, зокрема Алан Тьюринг. Результатом їхньої роботи стала розробка електромеханічної машини «Бомба» [22, p. 5]. Проте ці пристрої, попри свою значущість, ще не мали електронних запам'ятовувальних пристроїв у сучасному розумінні. Обчислювальні потужності навіть найдосконаліших комп'ютерів тієї епохи були незрівнянно меншими за можливості сучасних мобільних ґаджетів.

Важливим етапом стало створення ЕОМ «Colossus» — першої програмованої машини, що мала певні ознаки вбудованої пам'яті. До цього всі електронно-обчислювальні машини використовували перфокартки (зображення машин «Бомба» та «Colossus», а також перфокарт, що використовувалися до впровадження нової архітектури, див. у Додаток , Іл. 4).

Саме впровадження архітектури фон Неймана стало революційним проривом, що уможливив подальший розвиток обчислювальної техніки. Сучасні нейронні мережі можна вважати вищим щаблем еволюції цієї архітектури, оскільки їхня ефективність, а також сам процес навчання (training), критично залежать від швидкості доступу до пам'яті та можливості опрацювання величезних масивів даних безпосередньо в ній.

1.1 Визначення поняття “deepfake” та історія його виникнення

Розглянемо історію виникнення терміну *deepfake*. Як зазначає стаття у електронному блозі групи компаній «Norton Security» [16] комбінація слів deep learning (глибоке навчання) та fake (підробка). Широке поширення він отримав наприкінці 2017 року, коли на платформі Reddit з’явилася спільнота з однойменною назвою “Deepfakes”. Учасники цієї спільноти почали публікувати відео, в яких за допомогою алгоритмів машинного навчання замінювали обличчя людей , найчастіше для створення порнографічного контенту [13, p. 4]. З моменту своєї появи термін deepfake стрімко увійшов у публічний дискурс, особливо після того, як медіа почали активно висвітлювати потенційні загрози цієї технології, зокрема її використання для дезінформації та маніпулювання громадською думкою. Такі платформи як YouTube і Twitter були заповнені відео, створеними за допомогою deepfake, як з розважальною метою, так і з навмисно маніпулятивним змістом.

Отже, історія виникнення терміну deepfake нерозривно пов’язана з демократизацією доступу до технологій штучного інтелекту та розвитком генеративних змагальних мереж (GAN). Поява цього феномену ознаменувала перехід від складних візуальних ефектів, доступних лише великим кіностудіям, до інструментів, якими може скористатися пересічний користувач персонального комп’ютера. Це стало можливим завдяки відкритому вихідному коду та спільнотам ентузіастів, які, експериментуючи з алгоритмами машинного навчання, ненавмисно створили потужний інструмент для викривлення реальності. Таким

чином, термін швидко трансформувався з назви нішевої інтернет-субкультури у загальновживане поняття, що описує глобальний виклик інформаційній безпеці XXI століття.

Визначення цього явища прямо відповідає його технічним характеристикам, а саме з технічної точки зору, технологія deepfake ґрунтується на принципах глибокого навчання, зокрема на використанні згорткових нейронних мереж (CNN) та генеративно-змагальних мереж (GAN). Згорткові нейромережі застосовуються для аналізу й виділення ключових ознак у зображеннях або відео, тоді як генеративно-змагальні мережі відповідають за створення нового контенту [26, p. 5] , який максимально схожий на оригінал і складно відрізняється неозброєним оком.

Із точки зору соціології, соціальної взаємодії, deepfake можна охарактеризувати як певний візуальний продукт, який містить контенту частину, яка відрізняється від оригіналу і / або не відповідає по змісту оригінальному візуальному продукту яка відрізняється від оригіналу і / або не відповідає по змісту оригінальному візуальному продукту, проте сприймається аудиторією як автентична. У ширшому контексті, deepfake — це не лише візуальна маніпуляція, а комплексний феномен, що охоплює синтез або модифікацію будь-якого типу медіаконтенту [27 , p. 2] (аудіо, відео, зображення, текст) за допомогою методів штучного інтелекту з метою створення ілюзії реальності.

1.2 Стан наукової розробки теми deepfakes у сучасному дослідницькому просторі

Упродовж останніх років тема deepfakes отримала широке висвітлення в науковій, медійній і правозахисній спільнотах. Важливу роль у теоретичному осмисленні deepfake-технологій відіграють праці у сфері комп’ютерних наук, зокрема у галузі глибокого навчання, штучного інтелекту (AI), а також філософії симулякрів і постправди (Ж. Бодрійяр [9], Н. Бостром [10] й ін.). Дослідники активно вивчають як алгоритмічні принципи створення deepfakes (особливо

GAN — Generative Adversarial Networks), так і етичні, соціальні та правові наслідки їх поширення.

Налаштуванням та доробкою вже готових алгоритмів займаються розробники із Південної Кореї [7], зокрема Джихе Бек (Jihye Back), науковець із Сеульского Національного Університету (Seoul National University), один із алгоритмів, що був описаний цим науковцем зараз використовується у абсолютно прикладних речах – створенні deepfake зображень у стилі «cartoon» (з англ. – мультфільм, мультиплікаційний твір)

У таких джерелах як IEEE Transactions, наприклад, фундаментальний технічний труд Абхиджит Пуннапарта (Йоркський Університет, Торонто, Канада) [41], Єгор Захаров (Федеральний Інститут технології, Цюрих, Швейцарія)[52] . На працях останнього науковця хотілося б зосередити більшу увагу. Саме його технічні досягнення та матеріали значною мірою стали прикладами та наочними демонстраціями сучасних технічних здібностей технології deepfakes, що в свою чергу лягла у основу соціально-етичних аспектів цієї роботи. Ця технологія вже зараз допомагає вирішувати такі проблеми людства як генеративні нейронні мережі для спрощення обчислень вимірювання температури на адронному колайдері; створення цифрових «аватарів» людей у соціальних мережах, наприклад, як це зараз відбувається у Instagram; «нейронна стрижка», або моделювання волос на голові за допомогою нейромереж; генерування текстур для віртуальної або змішаної реальності.

У *N.Y.U. Journal of Legislation & Public Policy* фіксується стабільне зростання публікацій, присвячених як моделюванню та синтезу візуального контенту, так і детекції маніпуляцій за допомогою ШІ [31, p. 2]. Зокрема, активно досліджуються нейронні моделі «talking heads», відновлення raw-зображень, детекція фейкових облич, а також прикладна реалізація таких технологій у цифрових медіа.

У сучасному науковому дискурсі спостерігається зростаючий інтерес до феномену deepfake як складової ширшого процесу розвитку генеративного штучного інтелекту. Особливо активно ця тема розробляється у сфері цифрової

криміналістики. Одним із ключових фахівців у цій галузі є доктор Хані Фарід (Hany Farid), професор Каліфорнійського університету в Берклі, який має багаторічний досвід у сфері аналізу цифрових маніпуляцій та візуального шахрайства [20]. Він очолює аналітичну ініціативу, присвячену виявленню та документуванню випадків використання deepfakes у контексті президентських виборів у США 2024 року. На його спеціалізованій онлайн-платформі вже зафіксовано близько десятка резонансних прикладів deepfake-маніпуляцій, що активно циркулювали в американському інформаційному просторі, зокрема на платформах TikTok, Instagram та X (колишній Twitter). Низка провідних дослідників у галузі криміналістики здійснює системний моніторинг поширення deepfake-контенту в реальному часі.

Кількісний аналіз публікацій у відкритих наукових базах, здійснений наприкінці 2021 року, демонструє стрімке зростання академічного інтересу до тематики deepfakes. Хоча офіційно зафіксоване число рецензованих статей, можливо, не відображає повною мірою загальний обсяг наукових досліджень (через закритість деяких наукових платформ чи дублювання публікацій), статистична динаміка вказує на сталу тенденцію експоненціального зростання (див. Додаток , Іл. 11). Це свідчить не лише про розширення наукового поля теми, а й про міждисциплінарний інтерес до неї — з боку правознавства, медіастудій, філософії технологій, кібербезпеки, штучного інтелекту тощо.

З позиції філософського осмислення феномена deepfake та штучного інтелекту, ґрунтовний аналіз цієї теми запропонував український філософ і викладач Київського національного університету імені Тараса Шевченка професор Андрій Баумейстер. У низці публічних лекцій, розміщених у відкритому доступі, дослідник здійснює критичну рефлексію щодо впливу генеративних технологій на масову свідомість, проблеми втрати онтологічної стабільності в цифрову епоху, а також філософські наслідки розмиття межі між реальністю та симулякром у медійному просторі [54; 55]. Його підхід є важливим внеском у розуміння культурної та екзистенційної складової впливу deepfake-технологій на сучасне суспільство.

Із точки зору детекції діпфейків існує багато робіт та авторів, які займаються саме розробкою «захисту», а не «нападу» у сфері deepfakes, наприклад науковці університету Макао (Китайська Народна Республіка), Юхао Джу (Yuhao Zhu) та Кю Лі (Qi Li) із розробкою алгоритму MegaFS (Megapixel Face Swapping) , який може ідентифікувати діпфейки по одному зображенню із точністю у 95 % [53, p. 2–3]. З появою мультимодальних нейронних мереж, здатних перетворювати текст у зображення (Text-to-Image), процес створення deepfakes зазнав фундаментальних змін. Якщо ранні методи (такі як автоенкодери) вимагали маніпуляцій з візуальними даними на рівні пікселів та масок, то сучасні дифузійні моделі дозволяють керувати генерацією через природну мову. Цей процес, відомий як промт-інжиніринг (prompt engineering), став критично важливим інструментом для створення високоточних та контекстуально складних підробок, у цьому великих успіхів досягли науковці із університету Любляни, Словенія (University of Ljubljana) Даріан Томашевич (Darian Tomasevic) та Фаді Боутрос (Fadi Boutros) із інституту комп’ютерних графічних досліджень, Дармштадт, ФРН (Fraunhofer Institute for Computer Graphics Research IGD) [46, p. 4] і розробили унікальний алгоритм промптингу «ID-Booth».

Проблематика deepfakes вивчається в межах кількох дисциплін: комп’ютерних наук та інженерії (створення та оптимізація моделей); кібербезпеки (виявлення deepfake-контенту в режимі реального часу); медійних студій (вплив на суспільну свідомість і політичні наративи); правознавства (правове регулювання маніпулятивного контенту); етики та філософії технологій (поняття симулякрів, відповідальність за створення контенту). Водночас, попри активне дослідження, спостерігається нерівномірність розвитку теми: технічні аспекти розроблені глибше, тоді як правові та гуманітарні підходи лише формуються. Бракує уніфікованих механізмів оцінки загроз та єдиної системи нормативного регулювання на глобальному рівні. Хоча ця тема перекликається із юридичними аспектами функціонування deepfakes, які будуть описані у наступному розділі, але вже тут зазначу, що необхідні дії та дослідження приймаються, і один із таких

комплексних підходів – розробка суцільного та «всеосяжного» закону для штучного інтелекту та deepfakes як дочірньої технології.

Отже, аналіз наукового дискурсу засвідчує, що феномен deepfakes вийшов за межі суто технологічної проблеми, трансформувавшись у комплексний виклик для правової, правоохоронної та соціальної систем. Ключові наукові прогалини наразі зосереджені у площині відсутності уніфікованої методології оцінки ризиків, недостатнього вивчення механізмів комерціалізації кримінальних deepfake-послуг та критичної вразливості судової системи перед синтетичними доказами. Відтак, перспективи подальших досліджень полягають у переході від фрагментарного опису технології до створення міждисциплінарних моделей протидії, що об'єднуватимуть інструментарій цифрової криміналістики, правову регламентацію обліку синтетичного контенту та психологічні механізми захисту інформаційного простору від маніпулятивного впливу.

1.3 Основні теоретичні підходи до вивчення deepfakes: наукові прогалини та перспективи подальших міждисциплінарних досліджень

Незважаючи на динамічний розвиток тематики, залишається низка недосліджених питань: відсутність чіткої класифікації типів deepfakes за ризик-орієнтовними критеріями; обмежена кількість технологій детекції з відкритим кодом, які легко масштабуються; недостатнє вивчення впливу deepfakes на мікро-поведінковому рівні в соціальних мережах. У подальших дослідженнях доцільно інтегрувати підходи цифрової криміналістики, семіотики візуального контенту та когнітивної психології, що дозволяє глибше зрозуміти як механізми сприйняття, так і можливості контр наративів у середовищі інформаційної війни. Наприклад, вже зараз Європол (поліцейська європейська служба) зазначає, що «особи, які володіють складними навичками у сфері штучного інтелекту, можуть надавати такі послуги іншим суб'єктам, зокрема зловмисникам, які, не маючи технічної експертизи, здатні використовувати ці інструменти для складних атак

соціальної інженерії[1]» [19, p. 16]. Підкреслимо, що сьогодні недостатньою мірою досліджено етико-психологічний вплив при використанні deepfakes у медичних або терапевтичних цілях (як, наприклад, у deepfake-терапії).

Цей сегмент цифрової злочинності є малодослідженим у науковій літературі, що створює низку прогалин: відсутність повної типології платформ, які вже зараз починають пропонувати послуги deepfake на комерційній основі; низький рівень дослідження механізмів формування ринку таких послуг та їхньої інфраструктури; недостатність аналізу стратегії адаптації кримінальних суб'єктів до цих новітніх технологій; помилкової реакції поліції на фальсифіковані матеріали з соціальних мереж (наприклад, сфабриковані кадри з місця подій, таких як демонстрації); невірної ідентифікації підозрюваних через віртуально змодельовані кадри, що стають вірусними; дискредитації поліцейських шляхом створення підроблених матеріалів, що демонструють їхні фальшиві неправомірні дії.

Актуальною науковою проблемою є перегляд критеріїв автентичності аудіовізуальних доказів, що використовуються в судовому процесі. Deepfake-технології радикально ускладнюють завдання оцінки достовірності таких матеріалів, незалежно від їх джерела: мобільний пристрій, соціальні мережі, камери відеоспостереження. Це зумовлює потребу у розробці алгоритмів перевірки справжності цифрових матеріалів, побудові послідовного ланцюга зберігання цифрових доказів і створенні інституційних стандартів верифікації аудіовізуального контенту.

Отже, наукові прогалини та перспективи подальших досліджень полягають у необхідності системного переосмислення deepfake не лише як технологічної інновації, а як комплексного виклику для інформаційної та правової безпеки держави. Проведений аналіз засвідчує, що існуючий науково-методичний інструментарій не встигає за темпами адаптації кримінального середовища до можливостей генеративного штучного інтелекту. Критичною залишається відсутність уніфікованих стандартів верифікації цифрового контенту, що ставить

[1] Тут і далі цитати з англомовних джерел надаються в авторському перекладі.

під загрозу об’єктивність судочинства та ефективність правоохоронної діяльності. Відтак, подальші наукові розвідки мають зосередитися на формуванні міждисциплінарної парадигми протидії, яка б інтегрувала удосконалені алгоритми детекції, оновлені криміналістичні методики дослідження цифрових доказів та психологічні механізми захисту суспільства від дезінформаційного впливу.

1.4 Концептуальне розуміння штучного інтелекту та його ролі в розвитку deepfake-технологій

У сучасному науковому дискурсі термін deepfake (глибинна підробка) інтерпретується як результат роботи специфічних алгоритмів машинного навчання, здатних синтезувати медіаконтент із високим ступенем реалістичності. У вузькому технічному значенні — це методика синтезу зображень людини, що базується на комбінації існуючих візуальних даних для створення нових, штучних образів або накладання характеристик однієї особи на іншу.

Фундаментом цієї технології виступають глибокі нейронні мережі (Deep Neural Networks), зокрема дві ключові архітектури:

– Автоенкодери (Autoencoders): використовуються для стиснення вхідних даних (наприклад, обличчя) у латентний код і подальшого їх відновлення, що дозволяє «міняти місцями» ознаки різних облич.

– Генеративно-змагальні мережі (GANs – Generative Adversarial Networks): система, де одна нейромережа («генератор») створює контент, а інша («дискримінатор») намагається відрізнити його від реального. У процесі цього змагання якість підробки постійно зростає

Сучасні можливості deepfake-технологій еволюціонували від простої підміни обличчя (face-swapping) до комплексної симуляції особистості. Сьогодні алгоритми здатні синхронізувати рухи губ із довільним аудіотреком (lip-sync), клонувати голос із збереженням інтонаційних особливостей та імітувати унікальну мікро міміку людини. Цей прогрес став можливим завдяки конвергенції

досягнень у сферах комп'ютерного зору (Computer Vision) та обробки природної мови (NLP).

Втім, аналізуючи феномен deepfakes, необхідно критично підійти до самої дефініції «штучного інтелекту» (ШІ). Як дослідник даної проблематики, автор вважає за необхідне зазначити, що поширений термін «ШІ» є значною мірою маркетинговим конструктом, який не зовсім коректно відображає сутність технологічних процесів. Сучасні системи, які ми називаємо «інтелектуальними», фактично є складними статистичними інструментами, а не носіями розуму.

Для глибшого розуміння звернемося до класифікації, яку пропонує шведський філософ, професор Оксфордського університету Нік Бостром. Він виокремлює декілька градацій інтелекту:

– інтелект людського рівня (Human-Level Machine Intelligence): система, здатна вирішувати широкий спектр завдань, доступних людині, володіючи інтуїцією, здатністю до навчання та розумінням контексту;

– штучний інтелект (ШІ): у широкому сенсі — інтелект, створений у небіологічному середовищі;

– штучний суперінтелект: гіпотетична сутність, що радикально перевершує когнітивні можливості людини в усіх сферах [10, p. 4-5].

Але, як зазначають дослідники М. Альфонсека та співавтори [6, p. 4], традиційні підходи до етичного регулювання ШІ, такі як «Три закони робототехніки» А. Азімова, є недостатньо ефективними для супер інтелекту. Вони базуються на хибному припущенні, що програмісти здатні ідеально формалізувати етичні норми, а система не може їх автономно трансцендувати. Більш сучасні пропозиції, класифіковані Ніком Бостромом, включають методи «обмеження можливостей» (capability control), наприклад, ізоляція ШІ у локальному середовищі («boxing»), та методи «селекції мотивації» (motivation selection).

Погоджуючись із класифікацією Бострома, варто доповнити її філософським критерієм, сформульованим Рене Декартом: «Cogito ergo sum» («Я мислю, отже існую»). Справжній інтелект неможливий без суб'єктності — усвідомлення власного існування. У біології наявність самосвідомості перевіряють за

допомогою «дзеркального тесту» (mirror test) [10, p. 2]. Здатність впізнавати себе у відображенні притаманна лише вищим приматам, слонам, дельфінам та вороновим, що свідчить про наявність у них зачатків самосвідомості.

На противагу цьому, сучасні нейромережі, включаючи ті, що генерують deepfakes, не володіють ані самосвідомістю, ані здатністю до нестандартного вирішення проблем у людському розумінні. З технічної точки зору, так званий «ШІ» (наприклад, великі мовні моделі або генератори зображень) є ймовірнісними моделями. Це своєрідні «велетенські словники», або багатовимірні простори параметрів, де зв'язки між елементами (словами, пікселями) встановлюються шляхом аналізу мільярдів прикладів Алгоритм не «розуміє» запит, він лише передбачає найбільш вірогідний результат на основі попереднього навчання. Це процес патерн-матчингу (співставлення шаблонів), а не мислення. Єдина подібність до людського мозку — це архітектурний принцип зв'язку між штучними нейронами, однак природа обробки інформації залишається суто математичною та статистичною.

Саме ця «механістичність» та відсутність морального компасу в алгоритмах призводить до появи етично суперечливих явищ. Яскравим прикладом є комерціалізація технологій DeepNude . Як зауважує дослідник Педчено Олександр (Pedchenko Oleksandr) [38, с. 2] Останнім часом у месенджерах (зокрема Telegram) набули популярності боти, що пропонують послугу автоматичного «роздягання» людей на фотографіях. Алгоритм, навчений на тисячах зображень людського тіла, просто домальовує відсутні, на його «думку», елементи (наготу) замість одягу.

Для нейромережі це лише виконання математичної функції мінімізації помилки генерації пікселів. Проте у соціальному вимірі це перетворюється на інструмент гендерно-зумовленого насильства та порушення приватності. Незважаючи на спроби впровадження прихованих водяних знаків та етичних обмежень розробниками, попит на подібний контент демонструє, як потужна технологія, позбавлена справжнього інтелекту та розуміння наслідків, стає зброєю в руках користувачів.

Таким чином, розглядаючи deepfake-технології, ми маємо справу не з «розумними машинами», а з високоточними інструментами імітації, які вимагають ретельного правового та етичного регулювання саме через свою ефективність та доступність.

1.5 Алгоритміка та машинерія продукування deepfakes

Один із найбільш широко застосовуваних у сфері deepfake типів нейронних мереж — це генеративно-змагальні мережі (GAN). Такі моделі складаються з двох ключових компонентів: генератора і дискримінатора. Генератор відповідає за створення зображень на основі заданих вхідних параметрів, тоді як дискримінатор аналізує отримані зображення й оцінює, чи є вони справжніми або синтетичними. Завдяки постійній взаємодії та суперництву між цими двома елементами вдається досягати високореалістичного візуального результату.

Класична архітектура GAN включає дві нейромережі: генератор (G) і дискримінатор (D), які функціонують за принципом змагання. Генератор прагне створити зображення, які візуально не відрізняються від реальних, а дискримінатор намагається виявити, які зображення є фальшивими. Мета цієї «гри» — удосконалити генератор до такої міри, щоб дискримінатор більше не міг відрізнити підробку від оригіналу. Якість взаємодії між двома мережами оцінюють за допомогою функції виграшу V(D, G), яка слугує мірою ефективності генерації [15 p. 12]. Окрім базової архітектури GAN, значну роль у створенні deepfakes, особливо у задачах заміни обличчя (face swapping), відіграють автоенкодери (Autoencoders). Ця технологія базується на стисненні вхідного зображення обличчя в компактний латентний вектор (кодування) та його подальшій реконструкції (декодування). Для реалізації підміни використовується один спільний энкодер для двох осіб, але різні декодери. Це дозволяє «витягнути» міміку та кут повороту голови з вихідного відео (source) і накласти їх на цільове обличчя (target), використовуючи декодер, навчений саме на цільовій особі.

Подальший розвиток генеративних моделей призвів до появи архітектур типу StyleGAN, які дозволяють керувати процесом синтезу на безпрецедентному рівні деталізації. Ключовою особливістю таких мереж є використання механізму адаптивної нормалізації (AdaIN), який дозволяє розділяти атрибути зображення на «стилі» різного рівня. У Додатку Г подано візуальні приклади зображень, отриманих у результаті комбінування двох латентних кодів на трьох різних рівнях масштабування. Кожен із підпросторів латентного простору керує окремими високорівневими характеристиками зображення, що дає змогу точно контролювати стиль і структуру створеного візуального контенту: грубий рівень (Coarse styles) (відповідає за позу, форму голови та загальну геометрію обличчя); середній рівень (Middle styles): (визначає риси обличчя, форму очей, носа та зачіску); тонкий рівень (Fine styles): деталізує кольорову гаму, мікро текстуру шкіри та освітлення.

Загалом, саме можливість «змішування стилів» (Style Mixing) дозволяє створювати гіперреалістичні симулякри, де ідентичність однієї людини органічно поєднується з атрибутами іншої без видимих артефактів склеювання.

Висновки до розділу 1

У розділі здійснено комплексний аналіз феномена deepfakes як складного багатовимірного явища, що перебуває на перетині технологій штучного інтелекту, цифрових медіа, етики, права та соціальних наук. Ретроспективний огляд динаміки розвитку технології засвідчує фундаментальні трансформації: за вісім років з моменту появи у публічному просторі (2017 р.) deepfake еволюціонував від нішевих експериментів та технологічних курйозів до масового інструменту, щільно інтегрованого в сучасну цифрову реальність.

Ключовим фактором цих змін став стрімкий прогрес генеративних змагальних мереж (GAN) та розвиток методів one-shot learning. Це призвело до повної «демократизації» доступу: технічні бар'єри входу фактично нівельовано, а створення переконливого синтетичного контенту стало доступним пересічному

користувачеві за допомогою смартфона. Сучасні алгоритми усунули більшість технічних недоліків (артефактів, розсинхронізації), що досягло рівня фотореалістичності, який унеможливлює візуальну ідентифікацію підробки без спеціалізованого програмного забезпечення.

Встановлено, що науковий дискурс навколо deepfakes характеризується певною диспропорцією: технічні аспекти (синтез, детекція) опрацьовані значно глибше, ніж гуманітарно-правові виміри, які все ще перебувають у процесі формування. Залишаються суттєві наукові прогалини щодо класифікації ризиків, масштабування детекційних технологій та правових викликів у судочинстві, що актуалізує потребу в міждисциплінарному підході із залученням фахівців з кібербезпеки, права, психології та медіа студій.

Феномен демонструє яскраво виражену дуалістичну природу: володіючи значним потенціалом для інновацій у медицині, креативних індустріях та маркетингу, він водночас перетворився на потужну зброю в інформаційних війнах та інструмент високотехнологічного шахрайства. Це створює нові виклики для цифрової криміналістики та прав людини, формуючи об’єктивну потребу у розробці всеосяжного регуляторного поля та стандартизації процедур аутентифікації контенту.

Підсумовуючи, слід констатувати, що deepfake трансформувався у потужний соціотехнічний феномен. Його вплив виходить далеко за межі суто технічних питань обробки даних, фундаментально змінюючи парадигму сприйняття інформації, провокуючи кризу довіри до візуального контенту та змушуючи людство переосмислювати поняття автентичності, ідентичності та реальності в цифрову добу.

РОЗДІЛ 2
АЛГОРИТМІЧНІ ТА МОДЕЛЬНІ ОСНОВИ DEEPFAKES: АНАЛІЗ СУЧАСНИХ ПІДХОДІВ

Другий розділ роботи присвячено детальному аналізу технічного фундаменту технології deepfake. Розвиваючи теоретичні положення щодо генеративно-змагальних мереж (GAN), розглянуті у попередньому розділі, тут ми зосередимо увагу на конкретних архітектурах та алгоритмах, які роблять можливим створення синтетичного контенту. Зокрема, буде проаналізовано принципи роботи автоенкодерів та згорткових нейронних мереж (CNN), які є будівельними блоками сучасних систем генерації.

Крім теоретичних моделей, у розділі розглядаються прикладні методи маніпуляції медіа даними, такі як заміна обличчя (face swapping), модифікація міміки (face reenactment) та синтез мовлення. Метою цього розділу є не лише опис технічних інструментів, а й демонстрація того, як еволюція алгоритмів спростила процес створення фейків, зробивши його доступним для широкого загалу та перетворивши на значний виклик для інформаційної безпеки .

2.1 Огляд основних алгоритмів штучного інтелекту для створення deepfakes: автоенкодери, згорткові нейронні мережі, заміна обличчя (face swapping) , face reenactment (Модифікація міміки)

Найбільш небезпечними deepfake-технології виявляються під час виборчих кампаній, коли їх використання може серйозно вплинути на громадську думку, викликаючи сумніви навіть у відданих прихильників певного політичного діяча. Так, нещодавно одне з провідних американських медіа *Newsweek*, викрило deepfake із зображенням Дональда Трампа [18]. На змонтованому зображенні він начебто танцює з 13-річною дівчиною. До зображення також було додано текстовий супровід, який мав дискредитуючий характер: "Фотографія Дональда Трампа на приватному острові Епштейна, де він танцює з 13-річною дівчиною.

Йому тоді було 50 років. Як називають таких чоловіків?" [28]. У центрі уваги цього розділу — саме технічні аспекти створення deepfake-контенту, без яких неможливо зрозуміти механізм його функціонування. На сьогодні існує цілий набір інструментів і методів, що дозволяють генерувати відео, зображення чи аудіо з високим рівнем реалістичності, який складно відрізнити від справжнього.

Основу більшості сучасних систем генерації deepfake становлять генеративно-змагальні мережі (GAN), а також автокодери та згорткові нейронні мережі (CNN). Конкретно, яким саме чином працюють генеративно-змагальні мережі було пояснено мною у першому розділі цієї роботи, також, потрібно пояснити що таке автоенкодери та згорткові нейронні мережі.

Автоенкодери (Самокодуючі нейромережі). Це нейромережі, які вчаться стискати інформацію, а потім відновлювати її назад. Уявіть, що є зображення. Автоенкодер спочатку робить із неї спрощену версію (це називається кодування), а потім намагається відновити оригінал із цієї спрощеної версії (декодування). Якщо він робить це добре - значить, він зрозумів, що найголовніше в цій картинці. Використовуються, наприклад, щоб: прибирати шум із зображень, стискати дані, вивчати структуру даних.

Згорткові нейронні мережі (Convolutional Neural Networks, CNN). Це нейромережі, які «дуже добре» бачать картинки [15, p. 2]. Вони працюють так: спочатку беруть маленький «фільтр» і прокочують його по зображенню, щоб знайти важливі деталі - наприклад, кути, лінії, текстури. Потім передають цю інформацію далі, на інші рівні, де збираються складніші образи - наприклад, обличчя, очі, або посмішку. Використовуються, наприклад, щоб: розпізнавати обличчя, знаходити об'єкти на фото, розуміти, що зображено на зображенні.

Залежно від конкретного застосування, використовуються різні підходи: заміна обличчя на відео в реальному часі (face swapping), модифікація міміки (face reenactment), синтез мовлення (voice synthesis) чи повна генерація зображення на основі текстового опису (text-to-image generation). Демонстрація всіх вищеописаних технологій відображена у Додатуку, Іл 2. до цієї роботи.

Face Swapping (Заміна обличчя). Однією з ключових технологій, що лежать в основі створення deepfake-контенту, є face swapping — заміна обличчя на зображенні або відео із збереженням міміки, рухів та інших візуальних особливостей. Ця технологія активно використовується у кіноіндустрії, рекламі, відеоблогах, а також стала основою для багатьох deepfake-додатків. Один із таких додатків – Deep Live Cam [59], який дає змогу користувачу змінювати власне обличчя на обличчя іншої людини. Це програмне забезпечення може встановити і використовувати . Зображення того, як автор використовує технологію для того, аби замінити власне обличчя на обличча Ілона Маска відображено у Додатку, Іл 3. до цієї роботи.

Основні етапи роботи технології: 1) виявлення та сегментація обличчя , на першому етапі система за допомогою алгоритмів комп’ютерного зору (зазвичай на базі згорткових нейронних мереж) визначає присутність обличчя в кадрі та виділяє ключові точки — очі, рот, ніс, контури щелепи тощо. Це дозволяє точно окреслити область, яка підлягає заміні; 2) кодування особливостей обличчя-джерела , обличчя, яке буде перенесено на іншу людину, проходить через процес кодування. Використовуються автоенкодери або генеративні моделі, що стискають обличчя до латентного простору - математичного представлення, яке містить всі основні риси; 3) відстеження міміки та рухів голови – за допомогою алгоритмів трекінгу система аналізує, як змінюється обличчя оригінального відео: рухи губ, повороти голови, моргання, вираз емоцій. Ці дані потім використовуються для динамічного перенесення зображення нового обличчя; 4) генерація нового зображення або кадру відео — отримані латентні вектори обличчя-джерела накладаються на рухи та позиції обличчя-реципієнта. Генерація відбувається за допомогою моделей, таких як GAN (Generative Adversarial Networks), що дозволяє досягти високої фото реалістичності; 5) постобробка та фінальна інтеграція — на останньому етапі відбувається корекція кольорів, освітлення, контрасту, вирівнювання меж між зображеннями. Завдання — зробити результат максимально природним і непомітним для неозброєного ока.

Практичне використання цієї технології отримало широке розповсюдження не лише як розважальний інструмент, але й як компонент більш складних deepfake-систем. Програмні комплекси на кшталт DeepFaceLab, FaceSwap або Zao дозволяють виконувати заміну облич на відео в домашніх умовах, маючи базові навички роботи з ПК [30, p. 21]. Деякі з них навіть дозволяють працювати в реальному часі, що підвищує ризики зловживання технологією.

Face reenactment (Модифікація міміки). У квітні 2018 року медіакомпанія *BuzzFeed* оприлюднила відеоролик, створений за допомогою технології deepfake, у якому було зображено президента Барака Обаму. Метою цього відео було продемонструвати, наскільки просто можна змусити публічну особу "виголошувати" будь-який текст, який вона насправді ніколи не казала. У відео голос Обами звучав як справжній, що свідчить про ймовірне використання технології типу “puppet-master”, коли актор озвучує текст і відтворює міміку, яка згодом переноситься на обличчя іншої особи або ця технологія також відома як «модифікація міміки». Деякі фрази у відео були такими, що їх майже неможливо уявити у вустах справжнього Обами, що й посилювало ефект маніпуляції [58] . Як і було сказано у роботі раніше, сучасні deepfakes не працюють у якості «магічної кулі», натомість вони сприяють ерозії довіри.

Принцип роботи технології: 1) зчитування рухів обличчя джерела — камера або відеофайл фіксує рухи міміки джерела (source actor). За допомогою глибинного навчання визначаються ключові точки (landmarks) на обличчі: брови, повіки, губи, щелепа тощо; 2) моделювання тривимірної структури обличчя — алгоритм формує 3D-модель обличчя, на яку буде проектуватися міміка. Для цього використовуються нейромережеві архітектури, зокрема 3D Morphable Models (3DMM) або спеціальні моделі глибокого навчання (наприклад, базовані на GAN або ResNet); 3) перенесення виразів та рухів – отримана міміка переноситься на обличчя цільового персонажа (target identity). Цей етап потребує точного відображення нюансів руху — нахилів голови, посмішок, зморшок, моргань тощо; 4) Рендеринг фінального відео — після того, як всі зміни перенесено, генерується фінальне відео, в якому обличчя цільової особи поводиться так само, як і

оригінальне. Використовуються методи згладжування, корекції світла та тіні, щоб уникнути штучного вигляду. Європол у своєму звіті [19, с. 9] ще у 2024 році сформулював перелік рекомендацій щодо обмеження обігу синтетичного контенту, які необхідно імплементувати до національних законодавств країн ЄС. Проте на даний момент ці ініціативи залишаються переважно декларативними і перебувають на стадії стратегічного планування, не маючи імперативного характеру.

Отже, ми розглянули основні алгоритмів штучного інтелекту для створення deepfakes. Перейдемо до більш детального генеративних технологій створення дипфейків.

2.2 Voice Synthesis та Text-to-image generation

Технологію синтезу голосу (Voice Synthesis) доцільно розглядати у двох вимірах: як самостійний інструмент генеративного штучного інтелекту та як невід’ємну складову комплексних мультимодальних deepfake-систем. У контексті створення повноцінних цифрових двійників синтез мовлення найчастіше функціонує в тандемі з алгоритмами візуальної маніпуляції, такими як face swapping (заміна обличчя) та face reenactment (перенесення міміки) [51, p. 5]. Така інтеграція зумовлена необхідністю досягнення максимальної реалістичності: візуально досконалий образ вимагає відповідного аудіосупроводу, що імітує тембр, інтонацію та манеру мовлення об'єкта підробки. Використання синтезованого голосу дозволяє «озвучити» згенерованого персонажа, створюючи ілюзію повної присутності.

Водночас аудіодіпфейки (Audio Deepfakes) можуть виступати як автономний вид маніпулятивного контенту. Яскравим прикладом такого застосування є інцидент із поширенням у соціальних мережах сфальсифікованого запису голосу Джозефа Байдена, у якому політик нібито повідомляв про неминучий колапс фінансової системи [25]. Особливістю цієї підробки була відсутність візуального ряду: відео містило лише статичне чорне тло або нейтральну заставку. З технічної

та психологічної точки зору це був продуманий хід: імітація «злитого запису» поганої якості (шуми, перешкоди) не лише приховувала можливі дефекти синтезу, а й додавала контенту ознак документальності та автентичності, знижуючи поріг критичного сприйняття аудиторії.

Окремим, проте не менш важливим напрямом розвитку генеративних технологій, є моделі перетворення тексту в зображення (Text-to-Image generation). На відміну від алгоритмів заміни обличчя, які потребують вихідного фото- або відео матеріалів, ці системи (наприклад, DALL-E 3, інтегрований у ChatGPT, Midjourney, Stable Diffusion) створюють контент de novo — з нуля.

Загалом, процес генерації базується на інтерпретації нейромережею текстових запитів (prompts). Це породило нову сферу навичок — промт-інжиніринг, що полягає у вмінні формулювати точні інструкції для отримання бажаного візуального результату. На сьогодні технологія Text-to-Image активно інтегрується у медіавиробництво та маркетинг. Зокрема, автори YouTube-каналів та блогів масово використовують генеративні зображення для створення привабливих мініатюр (прев'ю) та візуального оформлення контенту. Це дозволяє створювати складні композиції, реалізація яких традиційними методами дизайну вимагала б значних часових та фінансових ресурсів (приклад використання такої deepfake-композиції для оформлення медіаконтенту наведено у Додаток, Іл. 1).

Висновки до розділу 2

Підсумовуючи викладений матеріал, можна констатувати, що технології deepfake вийшли за межі вузькоспеціалізованих технічних експериментів і стали невіддільною частиною повсякденного цифрового буття. Ключовою тенденцією сьогодення є тотальна «демократизація» доступу до нейромережевих інструментів.

Порівняльний аналіз із попередніми етапами досліджень демонструє кардинальну трансформацію ринку: якщо раніше генерування високоякісних

симулякрів вимагало специфічних технічних знань, значних обчислювальних потужностей або платних підписок, то сьогодні ці інструменти доступні широкому загалу безкоштовно та без необхідності реєстрації. Бар'єр входу у технологію фактично зник, що дозволяє створювати реалістичні аудіовізуальні підробки за допомогою простих текстових запитів (prompts) на побутовому рівні.

Водночас спостерігається значний розрив між швидкістю технологічної експансії deepfakes та реакцією правових інституцій. Хоча міжнародні організації визнають масштаб загрози, дієві механізми регулювання досі відсутні. Отже, цей розділ висвітлив технологічний та емпіричний рівень проблематики, продемонструвавши легкість створення та масштаби поширення deepfakes. Детальний аналіз нормативно-правового вакууму, який утворився навколо цієї технології, а також етичних викликів, що постають перед суспільством, буде здійснено у наступних розділах роботи.

РОЗДІЛ 3
СОЦІАЛЬНО-ЕТИЧНІ АСПЕКТИ ВИКОРИСТАННЯ DEEPFAKES

Єдиної, загальноприйнятої відповіді на питання, чим конкретно займається штучний інтелект, на сьогодні не існує. Майже кожен дослідник або автор, який пише про цю галузь, формулює власне визначення, виходячи з конкретного контексту та цілей дослідження. Саме тому поняття "штучного інтелекту" (ШІ) є багатогранним і охоплює цілу низку різнорідних напрямів і технологій.

У філософії досі немає єдиної думки щодо природи, сутності та статусу людського інтелекту. Як наслідок, не існує й чіткого критерію, за яким можна було б визнати комп’ютер "розумним". Саме це і було описано у вступі цієї роботи. Ще на ранніх етапах розвитку ШІ вчені пропонували певні концептуальні орієнтири - такі як тест Тюрінга, що мав на меті визначити здатність машини імітувати людську поведінку, або гіпотеза Ньюелла–Саймона [15], яка пов’язувала інтелект із можливістю вирішувати задачі та досягати цілей. Утім, жоден із цих підходів так і не став універсальним стандартом.

Попри концептуальну невизначеність, у науковій і технічній практиці склалися два головних підходи до створення систем штучного інтелекту, про які буде описано у наступному підрозділі.

3.1 Підходи до осмислення: низхідний підхід (Top-Down AI, семіотичний) та висхідний підхід (Bottom-Up AI, біологічний)

У сучасному науковому дискурсі концептуалізація штучного інтелекту (ШІ) традиційно розглядається крізь призму двох фундаментально відмінних підходів, які визначають архітектуру та логіку функціонування інтелектуальних систем: низхідного (Top-Down AI, або семіотичного) та висхідного (Bottom-Up AI, або біологічного). Низхідний підхід базується на спробі моделювати високорівневі когнітивні функції людини, такі як логічне мислення, мовлення, міркування, прийняття рішень, емоції та навіть творчість, шляхом маніпулювання символами.

Основними інструментами в цьому напрямку виступають експертні системи, бази знань та системи логічного виводу, що оперують формалізованою символічною інформацією. Мета цього напрямку полягає в тому, щоб навчити машину мислити подібно до людини, використовуючи чітко визначені абстрактні правила, логічні структури та онтології, як це описано в праці Філіпа Вербансіса та Джона Харгеса (Philip Verbancsics & Josh Harguess) із центру повітряних та морських бойових досліджень, Сан Дієго, США [49, p. 4]. Як зазначають дослідники , важливою метою для спільноти машинного навчання є створення підходів, здатних навчатися та знаходити рішення на рівні людських можливостей. Однією з областей, де люди тривалий час зберігали значну перевагу над алгоритмами, є візуальна обробка даних. Вагомим підходом до подолання цього розриву стали методи машинного навчання, що натхненні природними системами, такі як штучні нейронні мережі (ANN), еволюційні обчислення (EC), а також генеративні та розвиваючі системи (GDS). Дослідження у сфері глибокого навчання (Deep Learning) продемонстрували, що такі архітектури можуть досягати продуктивності, конкурентної з людською у певних візуальних завданнях, проте еволюція може відігравати значну роль у розвитку візуальних систем, що відкриває шлях до використання нейро еволюції (NeuroEvolution) у глибокому навчанні.

Діаметрально протилежним є висхідний підхід, який орієнтується на моделювання інтелектуальної поведінки «знизу догори», тобто через створення обчислювальних моделей, натхненних біологічними системами. Найбільш відомими прикладами реалізації цієї парадигми є штучні нейронні мережі, еволюційні алгоритми та нейроморфні обчислення. У центрі уваги дослідників цього напрямку перебуває вивчення того, як із простих біологічних елементів (нейронів) виникає складна емерджентна поведінка, та спроба відтворити ці процеси у вигляді комп’ютерних систем, таких як нейрокомп’ютери.

Слід зазначити, що останній підхід у класичному розумінні не завжди вважається частиною суто інженерної науки про штучний інтелект, а більше тяжіє до міждисциплінарного простору між інформатикою, біологією, нейронауками та когнітивною психологією. Проте саме завдяки конвергенції цих дисциплін

сучасний ШІ зробив найбільші практичні прориви, зокрема у сфері машинного навчання та генеративного моделювання. Найбільші досягнення ШІ останніх років, такі як ChatGPT, AlphaGo чи Midjourney, реалізовані саме завдяки нейромережам та генеративним алгоритмам, що базуються на висхідному підході. Яскравим прикладом еволюції в цьому напрямку є використання архітектури StyleGAN2 для задач перенесення стилю та генерації зображень. Як зазначає дослідниця Jihye Back, проблема збереження структурної цілісності вихідного зображення при генерації є критичною, і для її вирішення пропонується метод “файн тюнінгу” попередньо вивченої моделі, що дозволяє поєднувати структурні обмеження з ймовірнісним процесом генерації нових візуальних патернів [2, с. 3]. Такий підхід демонструє, як сучасні AI-системи долають обмеження жорсткої логіки, пропонуючи гнучкі інструменти для роботи з візуальним контентом.

Водночас для цілей даного дослідження видається більш пріоритетним комплексний аналіз того соціологічного стану речей, який демонструє сучасне західне суспільство, невід’ємною частиною якого є Україна. У попередніх наукових розвідках феномен deepfakes розглядався крізь призму класичних теорій Жана Бодріяра, Теодора Гейгера та Макса Вебера. Однак на сучасному етапі цей підхід потребує суттєвої актуалізації та переосмислення, оскільки кожна наукова дисципліна перебуває у стані безперервного розвитку. Класики соціологічної науки не стикалися з феноменом алгоритмічної генерації реальності, що ускладнює пряму екстраполяцію їхніх теорій на сьогодення без відповідної адаптації. Тому більш коректним методологічним кроком є аналіз сталих соціальних структур та їх трансформації під впливом новітніх технологій, що дозволяє розглядати deepfake не просто як технічну аномалію, а як інструмент соціального конструювання нової реальності.

Отже, аналіз підходів до осмислення штучного інтелекту демонструє фундаментальний зсув від символічних систем, що намагалися описати світ через правила, до конекціоністських моделей, які навчаються розуміти світ через дані. Саме домінування висхідного (біологічного) підходу уможливило появу технології deepfake, яка базується на здатності нейромереж виявляти та відтворювати

найтонші закономірності людської зовнішності та поведінки. Проте технічне розуміння алгоритмів є недостатнім для осягнення масштабу проблеми. Феномен deepfake вимагає виходу за межі інженерної площини у площину соціологічну, де він постає як каталізатор кризи довіри та інструмент переформатування соціальної реальності, що вимагає оновлення класичного соціологічного інструментарію.

3.2 Концептуалізація за Жаном Бодріяром (симулякри)

У своїй знаковій праці «Симулякри і симуляція» Жан Бодріяр розвиває концепцію гіперреальності, яка має глибоке значення не лише для філософії медіа, а й для соціології сучасного суспільства. Він наголошує, що в умовах постмодерного світу межі між реальністю та її представленням дедалі більше стираються. У соціологічному контексті це означає зростання ролі символічних структур, медіа та знаків у формуванні повсякденного соціального досвіду. Замість того, щоб відображати реальність, ці образи — симулякри — створюють нову соціальну реальність, яка живе за своїми законами.

Симулякр у трактуванні Бодріяра — це не просто копія без оригіналу, а структура соціального значення, яка виконує функцію заміщення реального у свідомості індивіда та колективу. Така реальність — гіперреальність — у соціальному сенсі стає ареною символічного контролю [9, p. 168], де істина більше не має стабільної основи, а сприйняття формується через знакові системи, нав’язані масовою культурою і технологічними медіа.

В умовах цифрової епохи, особливо з розвитком технології deepfake, ці ідеї Бодріяра набувають виняткової соціологічної актуальності. Deepfakes не просто спотворюють візуальну інформацію — вони впливають на соціальні уявлення про істину, довіру та автентичність. У соціології це можна розглядати як кризу легітимності символічних інститутів — медіа, політики, правосуддя, — які більше не можуть гарантувати достовірність інформації. Внаслідок цього формується новий соціальний ландшафт, де громадська думка дедалі частіше базується не на фактах, а на візуально переконливих симулякрах.

З точки зору соціології знання, deepfakes підривають епістемологічні основи сучасного суспільства, адже ускладнюють процес соціального конструювання реальності. Інформаційна взаємодія між індивідом та соціумом більше немає стабільного референту в об'єктивному світі. Це сприяє підвищенню соціальної тривожності, зростанню маніпулятивного потенціалу медіа і трансформації колективної ідентичності, яка дедалі більше визначається образами, створеними за допомогою алгоритмів, а не власним досвідом.

Отже, у контексті соціологічного аналізу, deepfakes є не просто технологічною новацією, а соціальним феноменом, що вимагає переосмислення таких понять, як "реальність", "довіра", "маніпуляція" та "ідентичність". Вони унаочнюють, як сучасні технології змінюють способи соціального сприйняття світу, впливають на структури влади і комунікації, та формують нові виклики для соціальних наук.

3.3 Етична проблематика deepfakes

Етична площина використання технологій deepfake характеризується значною складністю та неоднозначністю [34, с. 4]. Ключова дилема полягає у визначенні міри відповідальності між розробником інструменту та кінцевим користувачем. Хоча великі технологічні компанії дедалі частіше прописують жорсткі політики використання, це не стає гарантованим запобіжником від створення суперечливого контенту.

Часто розробники апелюють до концепції «технологічної нейтральності», стверджуючи, що програмний код сам по собі не є етичним чи неетичним — таким його робить намір людини. Наприклад, у дослідженні Consensus AI [45] підкреслюється проблематика того, що розробник часто займає позицію відстороненого спостерігача: користувач самостійно вирішує, для чого він буде використовувати згенерований контент, а платформа фактично знімає із себе етичні зобов'язання за наслідки.

У цьому контексті постає закономірне питання: наскільки взагалі етично розвивати та використовувати технології deepfakes? Безумовно, створення порнографії без згоди (так звана «порнопомста») є абсолютно неприйнятним та забороненим, зокрема політиками платформ, створених для розповсюдження контенту (детальніше див. Розділ 3 роботи). Однак, якщо виключити деструктивні сценарії, чи здатна ця технологія приносити суспільну користь?

Відповідь — так. Етичний аналіз вимагає неупередженого погляду на утилітарний потенціал технології. Центр Вудро Вільсона (Wilson Center) виділяє [37] низку позитивних аспектів застосування deepfakes, які вже сьогодні трансформують різні сфери життя: 1) подолання мовних бар'єрів та доступність контенту. Технологія синтетичного перекладу з можливістю синхронізації губ (lip-sync) відкриває нові горизонти в освіті та медіа. Це дозволяє адаптувати освітні лекції, новини чи кінематограф будь-якою мовою, зберігаючи візуальну органічність мовця. Така функція вже тестується та інтегрується в YouTube, що сприяє демократизації знань та глобальній комунікації; 2) покращення якості цифрової комунікації. Технології доповнення обличчя в реальному часі знаходять застосування у відеоконференціях. Яскравим прикладом є технологія Maxine від компанії Nvidia, розроблена для покращення зорового контакту та відгуку аудиторії. Алгоритм у режимі реального часу непомітно корегує напрямок погляду та міміку доповідача, створюючи ефект прямого зорового контакту з камерою, навіть якщо людина дивиться на екран або читає нотатки. Як зазначає сам контакт-центр компанії [60], технологія наразі перебуває в стані активного тестування та вдосконалення; 3) терапевтичне застосування в медицині. Одним із найбільш гуманних прикладів використання deepfakes є так звана «терапія спогадів» (reminiscence therapy). Для пацієнтів, які страждають на нейродегенеративні захворювання та вікові зміни, як-от хвороба Альцгеймера, можливо створювати спеціальні віртуальні симуляції. Як зазначає дослідниr Саар Хоек (Saar Hoek) із університету Амстердама (Нідерланди) [36, p. 3], використання образів молодих родичів або відтворення знайомих ситуацій з минулого допомагає пацієнтам заспокоїтися, активізувати пам'ять та покращити

емоційний стан. Дослідження свідчать, що це позитивно впливає на якість життя як самого хворого, так і його оточення, знижуючи рівень тривожності та агресії.

Як видно з цього короткого, але переконливого переліку, технологія deepfake має значний конструктивний потенціал. Вона не обмежується інструментами для шахрайства чи дезінформації, а є потужним засобом, що може бути використаний для гуманітарних, освітніх та медичних цілей. Отже, етичне завдання суспільства полягає не в забороні технології, а в створенні таких рамок її використання, які б мінімізували ризики зловживань, водночас не блокуючи інноваційний розвиток.

Висновки до розділу 3

Підсумовуючи аналіз соціальних та етичних аспектів функціонування deepfakes, можна констатувати, що ця технологія вийшла далеко за межі суто інженерної проблеми і трансформувалася у складний соціокультурний феномен.

У теоретичному вимірі розвиток deepfakes став результатом домінування «висхідного» (біологічного) підходу до штучного інтелекту, який через механізми машинного навчання дозволив створювати гіперреалістичні об'єкти. З точки зору соціології, спираючись на концепцію Жана Бодріяра, deepfakes виступають як досконалі симулякри — копії без оригіналу, що формують нову «гіперреальність». Це призводить до розмивання меж між істиною та вигадкою, провокуючи кризу довіри до аудіовізуальної інформації, яка століттями вважається об'єктивним доказом реальності.

Етичний аналіз демонструє глибоку амбівалентність цієї технології, яку доцільно розглядати крізь призму утилітарного підходу (співвідношення користі та шкоди для суспільства). З одного боку, як зазначають експерти Центру Вудро Вільсона, deepfakes відкривають революційні можливості в гуманітарній сфері: від подолання мовних бар'єрів через синтетичний переклад та покращення комунікації (Nvidia Maxine) до терапевтичного використання в медицині для пацієнтів із хворобою Альцгеймера.

З іншого боку, безконтрольне використання технології створює безпрецедентні загрози: від порушення приватності через створення нонконсенсуального порнографічного контенту до глобальних маніпуляцій громадською думкою в епоху «постправди». Ситуація ускладнюється позицією розробників (на прикладі Consensus AI), які часто декларують нейтральність інструменту, перекладаючи всю етичну відповідальність на кінцевого користувача.

Отже, людство опинилося в ситуації, коли традиційна етика, побудована на довірі до «обличчя» та візуального образу, виявляється не готовою до нових викликів. Це вимагає від суспільства не лише адаптації до факту існування deepfakes, а й вироблення нових механізмів верифікації правди та критичного осмислення реальності, де технологічний прогрес має бути збалансований чіткими етичними та, як буде розглянуто далі, правовими запобіжниками.

РОЗДІЛ 4
ПРАВОВІ ЗАСАДИ СТВОРЕННЯ ТА ПОШИРЕННЯ DEEPFAKES У МЕДІАПРОСТОРІ

Як і було описано раніше, регулювання deepfake в деяких державах і наддержавних формуваннях є або частковим, або відсутнє зовсім. Як буде зрозуміло з розділу цієї роботи, найчастіше deepfakes використовують або для створення «нонконсесуального» (без згоди) порнографічного контенту, або ж із метою шахрайства, і найчастіше, у кримінальній практиці йде пліч-о-пліч із шахрайством, здирництвом та іншими статтями подібного роду. Сама ж технологія deepfakes не є злочинною або ж караною, однак, у деяких країнах, таких як США, навмисне розповсюдження deepfakes заборонено під час передвиборчої кампанії в певних штатах. Бюрократія Євросоюзу лише створює нормативно-правові норми та рекомендації для країн-членів союзу. В Україні на цю тему зовсім не існує законодавчої або юридичної бази і поточний статус тематики deepfakes є поза юридичним полем.

4.1 Юридичні аспекти використання deepfakes: політика США

У Сполучених Штатах Америки наразі немає єдиного федерального закону, який би повсюдно та в усіх випадках забороняв або криміналізував створення та розповсюдження deepfake-контенту. Проте, на рівні окремих штатів, а також у вигляді законодавчих ініціатив на федеральному рівні, вже існують спроби врегулювати цю проблему, особливо коли йдеться про використання deepfake з шкідливою або злочинною метою. Зокрема, це стосується випадків, пов’язаних з переслідуванням осіб, створенням порнографії без згоди учасників або маніпулюванням виборчим процесом.

Серед найважливіших ініціатив можна виокремити наступні:

Законопроект про відповідальність за deepfakes (DEEPFAKES Accountability Act)[31] :

Цей законопроект був поданий до Палати представників США і мав на меті ввести кримінальну відповідальність за зловмисне створення та використання deepfake. Він передбачав обов’язкове маркування deepfake-контенту, а також покладав юридичну відповідальність на творців немаркованих deepfake у разі заподіяння шкоди. Станом на листопад 2023 року, цей законопроект ще не був прийнятий.

Акт про заборону зловмисних deepfakes (Malicious Deep Fake Prohibition Act) [33]: Інша ініціатива, вже в Сенаті США, передбачає криміналізацію зловмисного створення та розповсюдження deepfake-контенту. Хоча конкретні покарання ще не визначені, вони мали б бути встановлені після ухвалення закону. Наразі цей акт також перебуває на стадії розгляду.

Потрібно розуміти, що багато законів у США приймаються місцевими парламентами та штатами, тож і тут є деякі зміни у регулюванні deepfakes, а саме деякі штати вже ухвалили свої закони щодо deepfakes. Наприклад, у Каліфорнії діють два важливі закони:

– Законопроект Асамблеї № 602 дозволяє особам, які стали жертвами deepfake з сексуально експліцитним змістом, подати до суду, якщо такий контент був створений без їхньої згоди.

– Законопроект Асамблеї № 730 забороняє поширення аудіо- та відео-deepfake-матеріалів, які мають на меті дискредитувати кандидата на виборну посаду, за 60 днів до дати виборів.

Важливим етапом у формуванні реальної правозастосовної практики щодо синтетичного контенту став прецедент, створений Федеральною комісією зі зв'язку США (Federal Communications Commission — FCC) у вересні 2024 року. Цей кейс демонструє перехід від декларативних заборон до застосування жорстких фінансових санкцій за використання технологій клонування голосу у політичній боротьбі.

Предметом розгляду стало використання політичним консультантом Стівом Крамером технологій штучного інтелекту для створення аудіо-діпфейку (voice cloning), що імітував голос президента США Джо Байдена. За допомогою

автоматизованих дзвінків («рободзвінків») зловмисник поширював сфальсифіковане повідомлення, яке закликало виборців штату Нью-Гемпшир утриматися від участі у праймеріз.

У відповідь на цей інцидент FCC застосувала рекордні штрафні санкції, які базувалися на розширювальному тлумаченні існуючого законодавства. Зокрема, Комісія постановила, що використання технологій штучного інтелекту для генерації голосу в автоматичних дзвінках підпадає під регулювання Закону про захист прав споживачів телефонного зв'язку (Telephone Consumer Protection Act — TCPA) [23]. На підставі цього Стіва Крамера було оштрафовано на 6 мільйонів доларів США. Окрім покарання безпосереднього виконавця, цей кейс створив важливий прецедент відповідальності посередників. Телекомунікаційний провайдер Lingo Telecom, який технічно забезпечував маршрутизацію цих дзвінків, також був притягнутий до відповідальності та оштрафований на 1 мільйон доларів. Це сигналізує про те, що технічні посередники та платформи більше не можуть ігнорувати походження трафіку та контенту, який вони передають, особливо в контексті використання генеративного ШІ. Таким чином, даний випадок ілюструє, що правозастосовна система США здатна ефективно адаптувати існуючі нормативні акти (такі як TCPA) для протидії новітнім загрозам deepfake ще до прийняття спеціалізованого законодавства, покладаючи фінансову відповідальність як на організаторів атак, так і на технічних провайдерів.

Ключовим аспектом даного рішення стала правова кваліфікація дій порушника. Федеральна комісія зі зв'язку (FCC) застосувала норми Закону про захист прав споживачів телефонного зв'язку (Telephone Consumer Protection Act — TCPA). Регулятор офіційно роз'яснив, що дзвінки, згенеровані штучним інтелектом, підпадають під визначення «штучних або попередньо записаних голосів» (artificial or prerecorded voices), використання яких суворо регламентується цим законом, а здійснення автоматизованих дзвінків («рободзвінків») на житлові телефонні лінії без попередньої згоди абонентів є незаконним. Використання технології клонування голосу для створення синтетичного аудіо, що імітує реальну людину, не звільняє організатора від

відповідальності, а навпаки — є обтяжуючою обставиною. Крім того, у діях Стіва Крамера було зафіксовано використання технології спуфінгу (Caller ID spoofing) — навмисної підміни ідентифікатора абонента для введення отримувачів в оману щодо джерела дзвінка, що також є порушенням федеральних правил телекомунікацій. Сума штрафу у 6 мільйонів доларів сформувалася шляхом акумуляції штрафних санкцій за кожне окреме порушення (per-call fines).

Підсумовуючи, можна констатувати, що правове регулювання deepfake у США розвивається за двома паралельними векторами. З одного боку, триває процес формування спеціалізованого законодавства на федеральному рівні (DEEPFAKES Accountability Act, Malicious Deep Fake Prohibition Act), яке має на меті криміналізувати створення та поширення шкідливого синтетичного контенту. З іншого боку, як демонструє прецедент зі штрафом від FCC, державні регулятори вже зараз активно адаптують існуючі норми (зокрема, у сфері телекомунікацій та захисту прав споживачів) для реагування на новітні технологічні виклики. Такий підхід дозволяє заповнювати правові прогалини ще до прийняття комплексних законів, створюючи реальні механізми відповідальності не лише для безпосередніх виконавців атак, а й для технічних посередників, що сприяє формуванню більш безпечного інформаційного середовища.

4.2 Європейський регуляторний контекст

Що стосується європейського контексту, тобто країн так званого "старого світу", на сьогодні не існує єдиного, всеохопного нормативно-правового акту, який би прямо регулював процес створення та розповсюдження deepfake-контенту на рівні всього Європейського Союзу. Проте ця тема дедалі частіше стає об’єктом уваги як регуляторів, так і правоохоронних органів. У спільному звіті Європолу та Європейського парламенту під назвою «Боротьба із глибокими фейками у європейській політиці» зазначено, що нормативне середовище, пов’язане з deepfake у ЄС, являє собою складну систему, яка охоплює як положення на рівні основного (конституційного) права, так і механізми жорсткого (обов’язкового) та

м’якого (рекомендаційного) регулювання [12, с. 9]. Ця система діє як на загальноєвропейському рівні, так і на рівні окремих держав-членів ЄС.

Європол у своїх рекомендаціях пропонує встановити мінімальні вимоги для контролю за використанням deepfake-технологій [19, с. 4]. Зокрема, контент, створений за допомогою технологій deepfake, повинен обов’язково супроводжуватись чітким маркуванням, яке б вказувало на його штучне або маніпулятивне походження. Це дозволить користувачам усвідомлено сприймати інформацію та зменшить ризик введення в оману. Програмне забезпечення, яке дозволяє створювати deepfake-контент, правоохоронні органи вважають таким, що підпадає під категорію «високого ризику». Це пов’язано з потенційною загрозою, яку подібні технології можуть становити для основоположних прав і свобод людини, таких як право на приватність, недоторканність честі та гідності, а також право на достовірну інформацію. У майбутньому можна очікувати появу більш структурованих і суворих законодавчих ініціатив у цій сфері — зокрема, у межах загального регламенту про штучний інтелект (AI Act) [19, с. 13], який вже розробляється і містить положення, що стосуються і deepfake

Таким чином, європейський підхід до регулювання deepfakes ґрунтується на концепції оцінки ризиків та пріоритеті прозорості. Замість повної заборони технології, законодавчі ініціативи ЄС спрямовані на створення механізмів, що зобов’язують чітко ідентифікувати синтетичний контент, аби запобігти маніпуляціям громадською думкою. Це свідчить про намагання європейських інституцій знайти баланс між підтримкою технологічних інновацій та захистом інформаційної безпеки й фундаментальних прав громадян.

4.3 Правова ситуація deepfakes в Україні

Як уже зазначалося раніше, тема deepfake практично відсутня в українському законодавчому полі як окремий об'єкт регулювання. На сьогодні в Україні не існує жодного спеціалізованого нормативно-правового акту, який би прямо регулював, дозволяв або забороняв створення та поширення

deepfake-контенту. Іншими словами, це питання залишається поза межами прямого правового регулювання.

Однак, згідно з оцінкою кандидата юридичних наук, доцента Харківського національного університету внутрішніх справ К. В. Юртаєвої, відсутність спеціального закону не означає повну правову невизначеність. Дослідниця пропонує розглядати проблему крізь призму вже існуючих інститутів цивільного та кримінального права [5, с. 36]. Зокрема, оскільки спеціальні норми відсутні, регулювання використання образу особи у deepfake-відео має спиратися на загальні норми охорони немайнових прав, закріплені у статтях 307 та 308 Цивільного кодексу України. Ці статті визначають вимоги до використання зображення особи під час фото-, відео- та кінозйомок, а також подальшого розповсюдження таких матеріалів.

За загальним правилом, створення діпфейку вимагає згоди особи, чий образ використовується. Юртаєва К. В. зазначає, що вільне використання (без згоди) можливе лише у виняткових випадках, чітко визначених законодавством:

– якщо зйомки проводилися відкрито на заходах публічного характеру (мітинги, конференції тощо);

– якщо йдеться про історичних постатей (за умови відсутності комерційної мети);

– якщо контент створюється з навчальною, науковою метою;

– якщо метою створення є карикатура, пародія чи інший твір сатиричного жанру [5, с. 36–37].

Саме аспект «пародії» найчастіше використовується розробниками додатків для заміни обличчя (face swapping) як аргумент на користь легальності їхнього продукту. Наприклад, розробники популярного українського застосунку Reface App підкреслюють, що метою сервісу є «креатив і фан», а не створення дезінформації. Відповідно до Закону України «Про авторське право і суміжні права», створення пародій допускається без згоди автора, але з обов'язковим зазначенням джерела.

Проте дослідниця наголошує на важливому нюансі: навіть пародійний дипфейк не повинен порушувати інші немайнові права особи, зокрема:

– право на гідність та честь (ст. 297 ЦК України);

– право на недоторканність ділової репутації (ст. 299 ЦК України);– право на конфіденційність приватного життя (ст. 301 ЦК України);

– норми суспільної моралі (ст. 3 Закону України «Про захист суспільної моралі») [5, с. 37].

Порушення зазначених прав створює підстави для судового захисту. Однак на практиці публічні особи часто свідомо уникають позовів через ризик так званого «ефекту Стрейзанд», коли спроба юридично заборонити або видалити контент призводить до його ще більшого вірусного розповсюдження.

У площині кримінального права ситуація є складнішою та небезпечнішою. Статистика свідчить, що у 96% випадків технологія deepfake використовується для створення порнографічного контенту без згоди зображених осіб (переважно жінок). В Україні такі дії можуть бути кваліфіковані за статтею 301 Кримінального кодексу України («Ввезення, виготовлення, збут і розповсюдження порнографічних предметів») [5, с. 38–39]. Проте специфічна норма, яка б враховувала саме цифрову симуляцію реальної людини (deepfake pornography), у вітчизняному кодексі наразі відсутня, на відміну від законодавства деяких штатів США (наприклад, Нью-Йорка), де вже діють заборони на «діджиталізовані» порнографічні матеріали.

Додатковим вектором загроз є шахрайство. Відомим є прецедент 2019 року, коли зловмисники використали синтез голосу директора компанії для незаконного переказу коштів в особливо великих розмірах. Також зростає ризик створення відеофейків за участю публічних осіб для фінансових маніпуляцій або політичної дезінформації. Такі дії можуть підпадати під ознаки статті 190 КК України (Шахрайство), але доведення вини ускладнюється технічною природою злочину.

Підсумовуючи, К. В. Юртаєва підкреслює, що deepfake слід розглядати не лише як технологічну інновацію, а і як нову соціально-правову загрозу. Поточного

законодавства недостатньо для протидії масштабним ризикам, таким як маніпуляція виборчими процесами або гібридні атаки [5, с. 41].

З огляду на стрімкий розвиток генеративного штучного інтелекту, українське законодавство потребує комплексного оновлення, що має охоплювати:

– юридичне визначення поняття deepfake та його форм;

– встановлення чітких критеріїв допустимості використання зображення особи у синтетичному контенті;

– криміналізацію створення deepfake з метою дифамації, шантажу (зокрема порнопомсти) або шахрайства;

– регулювання захисту постраждалих через цивільно-правові та адміністративні механізми, адаптовані до кращих міжнародних практик (зокрема ЄС та США).

Отже, на сучасному етапі правове регулювання deepfake в Україні носить фрагментарний характер і спирається на загальні норми, що не враховують специфіку синтетичних медіа. Хоча чинні статті Цивільного та Кримінального кодексів дозволяють частково захищати права громадян у випадках порушення приватності чи створення порнографічного контенту, вони не охоплюють увесь спектр загроз, таких як політична дезінформація чи гібридні атаки. Тому нагальною потребою є гармонізація українського законодавства з міжнародними стандартами та розробка спеціалізованих правових механізмів, які б чітко визначали статус deepfake-контенту, встановлювали відповідальність за його зловмисне використання та забезпечували ефективний захист інформаційного простору держави.

4.4 Аналіз опосередкованого впливу технології deepfake на економіку України та аналіз економічних загроз .

Проведений аналіз фахової літератури [4, с. 310] свідчить про те, що інтеграція технологій генеративного штучного інтелекту у фінансову сферу створює системні ризики для глобальної екосистеми. Як зазначають

М. В. Рябокінь, Є. В. Котух та О. І. Папилев, технологія deepfake трансформувалася з інструменту дезінформації на потужну зброю кіберзлочинців, здатну підірвати фундаментальну основу фінансових відносин — довіру клієнтів [4, с. 311]. Статистичні дані підтверджують критичність ситуації: у 2023 році кількість інцидентів із використанням deepfake у секторі FinTech зросла на 700 % порівняно з попереднім роком.

На основі систематизації даних, наведених у дослідженні [4, с. 315], можна виокремити ключові вектори загроз для фінансового ринку: Крадіжка цифрової ідентичності та обхід біометрії. Зловмисники використовують генеративні нейромережі для створення синтетичних клонів клієнтів. Це дозволяє обходити сучасні системи ідентифікації (KYC) та авторизувати шахрайські транзакції [4, с. 314]. Маніпулювання фінансовими ринками. Як наголошують експерти (зокрема, з посиланням на А. Гупту та М. Гупту), deepfakes здатні спровокувати штучну волатильність активів [4, с. 311]. Поширення фальсифікованих відеозаяв керівників корпорацій або впливових інвесторів може спричинити панічні розпродажі акцій чи криптовалют, завдаючи збитків ще до верифікації інформації. Також дослідники вказують на ризики використання deepfakes для дискредитації керівництва фінансових установ. Один переконливий фейк, де CEO компанії нібито робить неетичні заяви, може миттєво зруйнувати репутацію бренду [4, с. 315]. Крім того, технологія автоматизує атаки соціальної інженерії, дозволяючи шахраям імітувати голос довірених осіб [4, с. 315] для отримання конфіденційних даних персоналу.

Критичною проблемою залишається низький рівень готовності ринку. Згідно з опитуванням, проведеним авторами статті, 67 % компаній FinTech-сектору використовують більше одного інструменту ідентифікації [4, с. 319], проте майже 70 % оцінюють свою готовність до протидії deepfake-шахрайству як низьку або дуже низьку [4, с. 320]. Для ефективного захисту фінансові інституції повинні переходити до багатокомпонентних стратегій [4, с. 321], які включають технологічні рішення (Liveness Detection, блокчейн) та постійний моніторинг інформаційного простору.

Яскравим прикладом етичного самообмеження є політика агрегатора контенту для дорослих Pornhub. Адміністрація ресурсу повністю заборонила розміщення матеріалів, створених за допомогою нейронних мереж, аргументуючи це порушенням принципу «консенсуальності» (згоди) [14]. Позиція платформи полягає в тому, що особа, чиє обличчя було змодельоване або перенесене на відео за допомогою AI, не надавала згоди на участь у такому контенті, що є грубим порушенням її прав.

Соціальні мережі, які є основними каналами поширення новин, також впроваджують активні заходи протидії. Платформа X (раніше Twitter) розробила стратегію боротьби з маніпулятивними медіа, що включає маркування контенту. У випадках, коли алгоритми виявляють ознаки синтетичної зміни зображення, відео чи аудіо, система додає спеціальне повідомлення, яке попереджає користувача про фальсифікацію. Більше того, перед спробою взаємодії з таким твітом (лайк, репост), користувач отримує додаткове попередження, що створює бар'єр для несвідомого поширення дезінформації. Як зазнає журналістка видання TechCrunch Сара Перез (Sarah Perez) , платформа надає посилання на верифіковані джерела для перевірки контексту [40]. У випадках, коли deepfake становить пряму загрозу безпеці людей, він підлягає повному видаленню. Подібні механізми виявлення та маркування маніпулятивного контенту впроваджуються і компанією Meta (Facebook), яка також перебуває на етапі активної реалізації політик щодо AI-генерованого контенту.

Отже, економічний вплив технології deepfake не обмежується лише прямими фінансовими збитками від шахрайських дій, а загрожує фундаменту цифрової економіки — довірі. Ерозія довіри до методів дистанційної верифікації та захищеності транзакцій може призвести до стагнації розвитку FinTech-сектору в Україні. Враховуючи виявлений критичний розрив між технологічними можливостями зловмисників та рівнем готовності бізнесу, нагальною потребою стає впровадження нових стандартів кібербезпеки. Захист національної економіки від синтетичних загроз вимагає переходу до проактивної моделі, що поєднує

передові алгоритмічні рішення з підвищенням кваліфікації персоналу фінансових установ.

Висновки до розділу 4

Підбиваючи підсумки аналізу правових засад функціонування технології deepfake, можна констатувати, що глобальне законодавче регулювання цієї сфери перебуває на етапі становлення. Як засвідчив проведений у розділі огляд, підходи до розв'язання проблеми варіюються від фрагментарних заборон на рівні окремих штатів у США до спроб системного регулювання через оцінку ризиків у Європейському Союзі. В Україні ж правова база залишається на стадії формування, спираючись наразі на загальні норми цивільного та кримінального законодавства, які не враховують специфіку синтетичного медіаконтенту.

Однак, зважаючи на інерційність державних законодавчих процесів, ключову роль у протидії поширенню деструктивних діпфейків на сьогодні відіграють самі технологічні платформи та соціальні мережі. Саме політика саморегулювання приватних компаній стає «першим ешелоном» оборони інформаційного простору.

Отже, ефективна модель протидії загрозам deepfake неможлива виключно силами державних інституцій. Вона вимагає гібридного підходу, де законодавство встановлює чіткі межі відповідальності (кримінальної та цивільної), а технологічні платформи забезпечують технічну реалізацію цих обмежень через модерацію та маркування. Для України вкрай важливо врахувати цей досвід, гармонізуючи власне законодавство з європейськими нормами та налагоджуючи співпрацю з адміністраціями глобальних соціальних мереж.

РОЗДІЛ 5
“STOP DEEPFAKES”: АВТОРСЬКЕ ЕМПІРИКО-ПРАКТИЧНЕ ДОСЛІДЖЕННЯ

У межах практичної частини кваліфікаційної роботи розроблено методологію емпіричного дослідження, яка передбачає моделювання процесу створення синтетичного медіаконтенту (deepfakes) та комплексний аналіз механізмів його розповсюдження в інформаційному просторі.

Актуальність цього етапу роботи зумовлена специфікою поточної геополітичної ситуації. В умовах повномасштабної збройної агресії Російської Федерації проти України національний інформаційний простір зазнає безпрецедентного тиску. Він стає ареною для реалізації гібридних загроз, де дезінформаційні кампанії та маніпулятивні технології використовуються широким колом ворожих суб'єктів як інструмент когнітивної війни.

Особливу загрозу становить поширення deepfake-відео, в яких використовуються згенеровані образи військовослужбовців Збройних Сил України. Такий контент, як правило, містить наративи про «неминучу поразку», «зраду командування» або «тотальну несправедливість», маючи на меті деморалізацію суспільства та Сил оборони. Небезпека подібних інформаційно-психологічних операцій (ІПСО) посилюється експлуатацією емоційної вразливості аудиторії та маніпуляцією «ефектом авторитету». Використання візуального образу військового (людини у формі) автоматично підвищує кредит довіри до повідомлення, змушуючи реципієнта сприймати сфальсифіковану інформацію як «інсайдерську правду» про реальний стан справ на фронті.

Варто зазначити, що сучасні технології генеративного штучного інтелекту досягли критично високого рівня якості. Синтетичні відео, створені за допомогою передових нейромережевих архітектур, характеризуються фотореалістичністю та високою точністю синхронізації міміки, що робить їх візуальну ідентифікацію надскладним завданням для пересічного користувача. Хоча на даному етапі існують певні технічні артефакти та маркери, які дозволяють виявити підробку

(детальний огляд методів детекції буде наведено у відповідному розділі роботи), технологія deepfake перебуває в стані перманентної еволюції. Швидкість удосконалення алгоритмів генерації випереджає розвиток інструментів захисту, що в найближчій перспективі може нівелювати ефективність візуальних методів розпізнавання.

5.1 Аналіз серії кейсів deepfakes в українському медіапросторі

В умовах повномасштабного вторгнення технологія deepfake трансформувалася з інструменту розваг на небезпечний елемент кіберзброї, спрямований на дестабілізацію внутрішньополітичної ситуації. Як зазначає М. О. Прищепа, найбільш показовим кейсом використання цієї технології в українському медіапросторі стала інформаційна атака, здійснена 16 березня 2022 року [3, с. 50]. Суть інциденту полягала у поширенні сфальсифікованого відеозвернення Президента України Володимира Зеленського, в якому він нібито закликав українських захисників скласти зброю та повернутися додому. Первинними каналами розповсюдження стали російські соціальні мережі («ВКонтакте») та Telegram-канали. За даними дослідника, поширення відео охопило значну аудиторію: лише в одному з великих Telegram-каналів («Инсайдер UA») підробка отримала понад 307 тисяч переглядів та понад 5 тисяч поширень за короткий проміжок часу [3, с. 50]. Аналіз технічної складової цього кейсу вказує на низьку якість виконання діпфейку. Було виявлено розбіжності у пропорціях голови та тіла, неприродну міміку та невідповідність голосових характеристик реальному голосу президента [3, с. 52]. Цю оцінку підтверджують і міжнародні експерти, зокрема Шаян Сардарізаде з BBC Monitoring, який охарактеризував цю спробу як одну з найгірших підробок, які він бачив [3, с. 52]. Важливим фактором нівелювання загрози стала превентивна комунікація з боку Головного управління розвідки МО України, яке попереджає про підготовку такої провокації ще на початку березня 2022 року [3, с. 53]. Дослідження впливу цього кейсу на суспільство за допомогою інструментарію Google Trends показало кореляцію між

появою фейку та різким зростанням пошукових запитів «deepfake» та «Zelensky» як в Україні, так і у світі саме в період 16-17 березня [3, с. 53]. Це свідчить про те, що, попри низьку якість контенту, подібні атаки суттєво впливають на інформаційне поле.

Для системного аналізу подібних загроз в українському медіапросторі М. О. Прищепа пропонує застосовувати когнітивний підхід, зокрема побудову нечітких когнітивних карт (Fuzzy Cognitive Maps). Ця методологія, запропонована Бартом Коско, дозволяє моделювати причинно-наслідкові зв'язки між різними факторами впливу (концептами), такими як «Сприймана достовірність підробки», «Недовіра до офіційних джерел» та «Інформаційно-психологічний вплив» [3, с. 56–57]. Використання таких моделей дозволяє прогнозувати сценарії розвитку інформаційних атак та оцінювати ризики для національної безпеки.

Окремої уваги заслуговує тенденція переходу deepfake-технологій від інструменту дезінформації до засобу вчинення тяжких фінансових злочинів. Показовим прикладом вразливості вітчизняних систем біометричної ідентифікації став інцидент, викритий Національною поліцією України у вересні 2025 року. Правоохоронцями було ліквідовано організоване злочинне угруповання, яке використовувало технологію deepfake для незаконного оформлення кредитів на громадян України [2]. Цей кейс є одним із перших масштабних підтверджень використання генеративного штучного інтелекту для обходу банківських верифікаційних алгоритмів у вітчизняній практиці.

Механізм реалізації злочинної схеми полягав у комбінуванні методів соціальної інженерії та високотехнологічної маніпуляції. Отримавши доступ до персональних даних громадян (фінансових номерів, паролів), зловмисники здійснювали несанкціонований вхід до мобільних застосунків банків та порталу державних послуг «Дія» через систему BankID. Ключовим етапом атаки ставало проходження процедури фото- та відео верифікації, яка є обов'язковою для підтвердження особистості при оформленні фінансових послуг. Для цього організатори схеми генерували deepfake-відео, накладаючи обличчя потерпілих на

власне зображення, що дозволяло успішно імітувати «живу» присутність клієнта та вводити в оману автоматизовані системи безпеки.

Масштаби даного інциденту свідчать про системний характер загрози: жертвами шахрайської схеми стали щонайменше 286 громадян, а загальна сума збитків перевищила 4 мільйони гривень. Викрадені кошти конвертувалися у криптовалюту для унеможливлення їх відстеження. Цей прецедент актуалізує необхідність перегляду існуючих протоколів біометричного захисту (Liveness Detection) в українському фінансовому секторі, оскільки стандартні методи перевірки виявилися недієвими проти якісно згенерованого синтетичного контенту.

Окрім безпекових ризиків вважаю за необхідне проаналізувати і найяскравіший кейс останнього часу для військової пропаганди, а конкретно — спроба використання генеративного відео для контрпропаганди (на прикладі ситуації у м. Покровськ) . Окремим, нетиповим кейсом, який демонструє етичні та репутаційні ризики імплементації технологій штучного інтелекту в військову комунікацію, став інцидент, зафіксований аналітичним ресурсом DeepState. Предметом аналізу стала публікація відеоматеріалу на сторінці 425-го окремого штурмового батальйону «Скала» у соціальній мережі X (Twitter).

Контекст інциденту полягав у спробі інформаційної протидії: у відповідь на поширення ворогом кадрів присутності в центральній частині міста Покровськ, на українському ресурсі з'явилося відео, яке мало на меті продемонструвати контроль ЗСУ над цією територією та спростувати наративи противника. Однак, як встановили OSINT-аналітики, відеоряд мав ознаки штучної генерації (deepfake/AI-generated video). Ключовим маркером підробки стали візуальні артефакти, характерні для недосконалості сучасних генеративних моделей при обробці динамічних сцен. Зокрема, було зафіксовано аномальну поведінку об'єктів фону — так званий «танцюючий» знак пішохідного переходу (див. Додаток, Іл. 6 до цієї роботи). Такі темпоральні неузгодженості є типовими технічними помилками нейромереж, які не здатні стабільно утримувати геометрію об'єктів у кадрі протягом тривалого часу. Подібне мерехтіння та пересвітлення найчастіша

проблема сучасної генерації відео, про це пишуть розробники китайської компанії Тенсент (Tencent) Ху Чен (Xu Chen) та Дзюнвей Джу (Junwei Jhu) [11]. Цей кейс є важливим для дослідження з точки зору негативного впливу на довіру. Як зазначається у звіті аналітиків, використання фейкового контенту, навіть з метою підняття бойового духу або заспокоєння суспільства («медійні заяви про зачистку»), призводить до зворотного ефекту — ерозії довіри до офіційних джерел інформації.

Іншим подібним кейсом є спроба дестабілізації суспільства через фабрикацію заяв урядовців (кейс Ольги Стефанішиної). Черговим доказом систематичного використання технології deepfake російськими спецслужбами для розхитування соціально-політичної ситуації в Україні стала інформаційна атака, спрямована проти Віцепрем'єр-міністерки з питань європейської та євроатлантичної інтеграції Ольги Стефанішиної.

Як зафіксував Центр протидії дезінформації (ЦПД), у ворожих Telegram-каналах масово поширювалося сфальсифіковане відео, на якому посадовиця нібито заявляє про намір законодавчо заборонити виїзд за кордон жінкам віком від 18 до 27 років під проводом «збереження генофонду» [57]. Технічний аналіз відеоматеріалу вказує на використання технології клонування голосу та ліп сінку (lip-sync), накладених на реальне архівне відео інтерв'ю. Метою цієї операції було створення штучного суспільного напруження, провокування панічних настроїв серед жінок та дискредитація органів державної влади. Спростування фейку відбулося оперативно: факт підробки підтвердила сама Віцепрем'єр-міністерка, а ЦПД опублікував детальний розбір маніпуляції, зазначивши відсутність будь-яких подібних законодавчих ініціатив у Верховній Раді (див. Додаток , Іл. 7 до цієї роботи). Цей кейс демонструє, що deepfake-атаки стають дедалі більш таргетованими, обираючи чутливі соціальні теми для маніпуляцій.

Іншим кейсом використання нових технологій для військової пропаганди та соціальної дестабілізації може слугувати створення синтетичних спікерів для впливу на міжнародну аудиторію (справа «Сабіни Мелс»). Окремим вектором еволюції deepfake-загроз є використання технології для створення неіснуючих «лідерів думок» або інсайдерів, що дозволяє ворожій пропаганді легалізувати дезінформаційні наративи в інформаційному полі країн-партнерів. Яскравим прикладом такої тактики стала масштабна пропагандистська кампанія РФ у Німеччині.

Суть операції полягала у поширенні дискредитуючого фейку про те, що Президент України нібито придбав віллу, яка раніше належала Йозефу Геббельсу. Першоджерелом цієї новини виступила нібито колишня співробітниця німецької компанії BIM — «Сабіна Мелс». Аналіз ЦПД виявив, що особа «Сабіни Мелс» є повністю фейковою. Технічний механізм фальсифікації полягав у крадіжці цифрової особистості (digital identity theft): за основу візуального образу було взято фотографії реальної людини — американської блогерки Аліси Гедсон (Alyssa Gadson). За допомогою нейромережевих алгоритмів статичне фото було анімоване та озвучене [56]. Цей кейс (див. Додаток, Іл. 9) демонструє високий рівень небезпеки, оскільки технологія deepfake тут використовується для створення «достовірного джерела» (proxy source), що ускладнює верифікацію контенту для іноземної аудиторії. Кейс «Сабіни Мелс» ілюструє небезпечну трансформацію deepfake-технологій: перехід від імітації відомих публічних осіб до конструювання повністю синтетичних «агентів впливу» або псевдоінсайдерів. Така стратегія дозволяє зловмисникам вирішувати проблему довіри до джерела інформації, створюючи ілюзію об'єктивного свідка або експерта, який не має репутаційного шлейфу.

Ця тенденція набуває глобального масштабу і не обмежується лише російською пропагандою. Підтвердженням системності цього явища є викриття прокитайської дезінформаційної мережі, відомої як «Wolf News». Як зазначають А. Сатаріано та П. Мозур у своєму розслідуванні [43] (2023 рік), це був перший зафіксований випадок використання deepfake-технологій у державній

інформаційній кампанії, де новини озвучували нереальні люди, а комп'ютерно згенеровані аватари. Використовуючи доступне комерційне програмне забезпечення (зокрема платформи типу *Synthesia*), зловмисники створили фіктивних ведучих, які транслювали антиамериканські наративи та пропагували інтереси Комуністичної партії Китаю.

Критична небезпека даного підходу полягає у двох аспектах. По-перше, це економічна доступність та масштабованість: створення професійного відеоконтенту з «живим» ведучим тепер коштує лічені долари та займає хвилини, що дозволяє генерувати дезінформацію у промислових масштабах. По-друге, це фундаментальна складність верифікації. Якщо у випадку з deepfake-відео президента (наприклад, В. Зеленського) фактчекери можуть порівняти підробку з оригіналом голосу чи міміки відомої особи, то у випадку з «синтетичними особистостями» (як у Wolf News чи Сабіни Мелс) аудиторія немає еталону для порівняння.

Таким чином, генеративний ШІ стає інструментом не лише технічної фальсифікації, а й створення розгалужених мереж фіктивних медіа та першоджерел, де межа між фактом і фікцією стирається на рівні самої особистості комунікатора.

5.2 ШІ та deepfakes на службі високотехнологічного міжнародного криміналітету

Фундаментальним викликом сучасної кібербезпеки є так звана «дилема подвійного використання» (dual-use dilemma) штучного інтелекту. Ті самі алгоритмічні потужності, що дозволяють захисникам оптимізувати процеси реагування, стають зброєю в руках зловмисників. Якщо технологія deepfake, що базується на архітектурі генеративно-змагальних мереж (GAN), загрожує візуальний та аудіальний довірі, то зловмисні великі мовні моделі (LLM) трансформують ландшафт текстових та програмних загроз.

Як зазначають дослідники «Unit 24» [48], що межа між доброчесним дослідницьким інструментом та потужним генератором загроз стає дедалі тоншою, часто залежачи лише від намірів розробника та наявності етичних запобіжників. Спеціалізовані зловмисні LLM (Malicious LLMs), такі як WormGPT або KawaiiGPT, навмисно розробляються або модифікуються для обходу етичних фільтрів, пропонуючи кіберзлочинцям функціонал для автоматизації атак

Впровадження цих технологій призвело до демократизації кіберзлочинності, знизивши поріг входження для зловмисників. Основні наслідки цього процесу, як зазначають дослідники кібербезпекової компанії «Unit24» включають: 1) масштабування замість навичок (Scale over skill). Інструменти дають можливість низькокваліфікованим акторам (script kiddies) запускати складні кампанії, які якісно перевершують атаки минулого; 2) стиснення часу (Time compression). Життєвий цикл атаки, що раніше вимагав днів ручної праці, скорочується до хвилин, необхідних для введення правильного запиту [48]. В основі цих систем лежать складні нейронні мережі, здатні обробляти та генерувати інформацію на рівні, що наближається до людського.

На практиці зловмисні моделі, як вказує журналіст Раві Лакшманан (Ravie Lakshmanan) із видання «The hacker news» [29] демонструють дві критичні переваги: лінгвістична точність (моделі здатні генерувати текст, який є граматично бездоганним та психологічно маніпулятивним; це виводить соціальну інженерію (фішинг, BEC-атаки) на новий рівень, дозволяючи створювати переконливі повідомлення, що імітують стиль спілкування керівників компаній або довірених партнерів, уникаючи традиційних маркерів фішингу, таких як помилки у синтаксисі) і вільне володіння кодом (зловмисні LLM здатні швидко генерувати, налагоджувати та модифікувати функціональний шкідливий код). Яскравим прикладом комерціалізації цієї загрози є модель WormGPT. Побудована на базі відкрикої моделі GPT-J 6B, вона була донавчена на конфіденційних наборах даних, пов'язаних із шкідливим програмним забезпеченням [48]. Ця модель активно рекламувалася на хакерських форумах як «безцензурна» альтернатива легальним

сервісам, пропонуючи зловмисникам можливості створення фішингових листів та генерації коду шкідливих програм.

Еволюція загрози простежується у появі таких інструментів, як WormGPT 4 та KawaiiGPT. Останній, будучи безкоштовним та доступним інструментом, дозволяє навіть новачкам генерувати скрипти для горизонтального переміщення мережею (lateral movement) та ексфільтрації даних. Тестування показало здатність цих моделей створювати функціональні скрипти-вимагачі (ransomware) із вбудованими механізмами шифрування та автоматичною генерацією листів із вимогою викупу, що використовують тактику залякування та терміновості.

Отже, поява необмежених LLM разом із технологіями deepfake встановлює новий базовий рівень цифрових ризиків, де здатність миттєво генерувати повний ланцюжок атаки — від переконливого візуального образу до шкідливого коду — стає доступною широкому колу зловмисників.

5.3 Актуальне застосування технології deepfake для продажу цифрових товарів та послуг

Еволюція технологій генеративного штучного інтелекту призвела до виникнення абсолютно нового сегмента ринку цифрових послуг. Якщо на початкових етапах алгоритми заміни обличчя (face swapping) та стилізації зображень були інструментами розваг або технічних експериментів, то сьогодні вони стали основою для прибуткових бізнес-моделей.

Аналіз сучасного українського медіапростору, зокрема соціальних мереж та месенджерів (Telegram, Instagram), демонструє стрімке зростання пропозиції послуг, що базуються на технологіях нейромережевої обробки зображень. Яскравим прикладом такої комерціалізації є сервіси створення стилізованих портретів, «артів» на полотні або цифрових аватарів.

Як видно з рекламного матеріалу, розповсюдженого в мережі Telegram (див. Додаток, Іл. 5 до цієї роботи), користувачам пропонується послуга створення «романтичного портрета» на основі звичайної фотографії. Хоча в маркетинговій

комунікації використовується термінологія традиційного мистецтва («малюють на полотні»), технічна реалізація цього процесу базується на використанні алгоритмів Image-to-Image (трансформація зображення в зображення) та Face Swapping (перенесення рис обличчя).

Механізм надання такої послуги виглядає наступним чином:

– Вхідні дані: Клієнт надсилає фотографію людини (донора).

– Генерація: За допомогою нейромережевих моделей (наприклад, на базі архітектури Stable Diffusion з використанням розширень ControlNet або InsightFace) обличчя клієнта переноситься на заздалегідь підготовлений художній шаблон (prompt-based template). Алгоритм адаптує освітлення, текстуру шкіри та стилістику під «живопис» або «цифровий арт».

– Виробництво: Отримане цифрове зображення друкується на фізичному носії (полотні).

Цей кейс демонструє ключову перевагу використання deepfake-технологій у бізнесі — автоматизацію творчого процесу. Те, що раніше вимагало від художника десятків годин ручної праці, нейромережа виконує за лічені хвилини (у рекламі зазначено можливість «розрахувати ціну за 1 хвилину», що опосередковано вказує на повну автоматизацію процесу оцінки та генерації).

З економічної точки зору, це дозволяє суттєво знизити собівартість продукції та зробити персоналізоване мистецтво масовим товарам. Однак, з етичної точки зору, виникає питання прозорості: споживач не завжди усвідомлює, що купує не роботу художника, а результат алгоритмічної генерації.

Таким чином, технологія deepfake вийшла за межі створення фейкових новин чи розважального контенту і стала потужним драйвером для ринку електронної комерції, дозволяючи монетизувати потребу людей у персоналізації та естетичному самовираженні. Останнім часом мережею ширяться так звані «промпти» для створення фотографій за допомогою нейромережі. Для того, щоби користувачеві створити deepfake достатньо лише відкрити користувацький інтерфейс і ввести необхідний вже затовлений текст (темпліт) , як це виглядає на Додатку, Іл. 9 до цієї роботи.

5.4 Авторський досвід зі створення deepfakes

У своїй роботі я використав існуючі напрацювання, а саме одну з найбільш просунутих нейронних мереж для створення deepfakes з використанням алгоритму PuLID (Pure and Lightning ID customization). Авторами цієї архітектури є дослідники Зінан Гуо (Zinan Guo) та Янзе Ву (Yanze Wu), а інформацію про роботу моделі було отримано з їхньої публікації до 38-ї конференції з нейронних систем та обробки даних (38th Conference on Neural Information Processing Systems) [24]. Алгоритм, за допомогою якого була виконана практична частина цієї роботи, має певні ключові відмінності від традиційних методів персоналізації. Насамперед, він реалізує підхід «tuning-free», що дозволяє вбудовувати ідентичність особи без необхідності тривалого донавчання моделі під кожне окреме зображення.

Головною архітектурною інновацією є впровадження додаткової гілки навчання — Lightning T2I branch, яка інтегрована зі стандартним дифузійним процесом. Це дозволяє використовувати механізм контрастного вирівнювання (Contrastive Alignment): система одночасно обчислює шлях генерації з впровадженням ID та без нього, мінімізуючи різницю в реакції нейромережі на текстові підказки. Завдяки цьому PuLID вирішує фундаментальну проблему попередніх рішень (таких як IP-Adapter або InstantID) — він забезпечує високу точність відтворення рис обличчя (ID fidelity), залишаючи при цьому незмінними стиль [24, ст. 4], освітлення та композицію вихідного зображення, що є критично важливим для створення переконливих та високоякісних deepfakes. Кінцеве зображення, що було створено за допомогою метода PuLID, доступно у Додаток, Іл. 10 до цієї роботи. На ньому чітко прослідковується збереження текстури шкіри та коректна адаптація освітлення, що підтверджує високу ефективність обраного алгоритму для генерації переконливих симулякрів.

Отже, практична апробація алгоритму PuLID у межах даного дослідження засвідчила фундаментальну зміну в парадигмі створення синтетичного контенту.

Перехід від ресурсомістких методів, що вимагали тривалого тренування на великих масивах даних (tuning-based), до миттєвої генерації (tuning-free) із використанням механізмів контрастного вирівнювання дозволяє досягати фотореалістичних результатів за лічені секунди.

Наступним етапом авторського емпіричного дослідження стала верифікація доступності та ефективності новітніх інструментів генеративного штучного інтелекту. Об’єктом аналізу було обрано передові моделі компанії Google, зокрема відеогенеративну модель Veo3 (у технічній документації та інтерфейсі також фігурує як NanoBanana). Важливою особливістю сучасного етапу розвитку цих технологій є їхня інтеграція у загальнодоступні користувацькі інтерфейси (наприклад, через екосистему Gemini). Це свідчить про радикальне зниження порогу входження: створення високоякісного синтетичного контенту більше не потребує навичок програмування чи потужного локального обладнання, а зводиться до вміння формулювати текстові запити.

Процес генерації розпочався з етапу промт-інжинірингу. Як продемонстровано на Іл. 12 Додатку, взаємодія з нейромережею NanoBanana (Veo 3) відбувається через інтуїтивно зрозумілий діалоговий інтерфейс. Автор сформулював деталізований опис сцени, освітлення та динаміки руху, що дозволило моделі згенерувати відеоряд, який відповідає заданим параметрам. Результат генерації (Додаток, Іл. 13) демонструє здатність алгоритму підтримувати темпоральну узгодженість кадрів (temporal consistency), уникаючи типових для ранніх діпфейків артефактів мерехтіння, що робить відеоконтент придатним для використання в інформаційних кампаніях.

Окремим напрямом експерименту стало дослідження можливостей модифікації біометричних характеристик особи на статичних зображеннях. Використовуючи алгоритми deepfake, автором було реалізовано завдання gender-swap (зміна гендерної приналежності). Як видно на Додаток, Іл. 13 (друге зображення), нейромережа успішно трансформувала вторинні статеві ознаки (структуру волосся, риси обличчя, макіяж), зберігши при цьому впізнаваність базових антропометричних рис об’єкта. Це підтверджує небезпеку використання

подібних технологій для створення фейкових особистостей (sock puppets), які базуються на реальних прототипах.

Для валідації реалістичності отриманого синтетичного образу було проведено польовий експеримент у середовищі соціальної мережі Telegram. Згенерований жіночий портрет було використано як фото профілю (аватар) у популярному чат-боті для знайомств «Дайвінчік». Під час реєстрації автор вказав жіночу стать та відповідні вікові параметри (Додаток, Іл. 14). Метою цього етапу було перевірити так званий «тест Тюрінга» у візуальному форматі: чи зможуть пересічні користувачі відрізнити синтетичне обличчя від реального фото в умовах звичайної соціальної взаємодії.

Результати експерименту виявилися показовими. Профіль із deepfake-зображенням (після мінімальної кольорокорекції для надання природності) успішно пройшов модерацію платформи та за короткий проміжок часу отримав значну кількість позитивних реакцій («вподобань») від реальних користувачів-чоловіків (Додаток, Іл. 15). Відсутність сумнівів у аудиторії підтверджує гіпотезу про те, що сучасні deepfake-технології досягли рівня, достатнього для введення в оману непідготовленого спостерігача. Експеримент було зупинено на етапі отримання фідбеку задля дотримання норм цифрової етики та уникнення нанесення емоційної шкоди користувачам через активний обман (active deception).

Отже, експеримент підтвердив, що сучасні архітектури здатні нівелювати типові артефакти попередніх поколінь deepfakes, такі як невідповідність освітлення або втрата стилістичних деталей фону. Це свідчить про критичне зниження технічного порогу входу для створення високоякісних підробок, що робить технологію доступною широкому загалу та, відповідно, значно ускладнює процес візуальної верифікації контенту медіа фахівцями.

5.5 “Я розпізнати deepfakes?”: розробка практичних рекомендацій для медіа фахівців

Стрімка еволюція генеративного штучного інтелекту, зокрема дифузійних моделей, GAN (Generative Adversarial Networks) та масштабних систем синтезу мовлення (text-to-speech), призвела до драматичного підвищення реалістичності синтетичного медіаконтенту. Як зауважує дослідник Ган Пей (Gan Pei) зі Східнокитайського педагогічного університету (East China Normal University), у цьому контексті навички виявлення діпфейків стають критично важливими для журналістів, фактчекерів та інших медіапрацівників [39 p. 5]. Даний підрозділ узагальнює сучасний стан досліджень та пропонує алгоритм дій для медіаорганізацій, що базується на поєднанні автоматизованих інструментів та експертного аналізу.

Медіафахівці повинні усвідомлювати, що покладання виключно на автоматизовані класифікатори несе високі ризики. Більшість сучасних детекторів (SOTA) демонструють високу точність на еталонних наборах даних (наприклад, FaceForensics++), проте їх ефективність різко знижується в реальних умовах («in-the-wild»). Основними причинами відмов є:

– Зсув домену (Domain shift): Реальний контент часто містить артефакти стиснення, шуми передачі даних або нові методи маніпуляції, які не були присутні у тренувальних даних алгоритмів [12, p. 2–3]. Наприклад, детектор навчався розпізнавати підробки на високоякісних відео, знятих у студії. Проте у реальному житті медіа фахівці працюють із відео, які користувачі знімають на звичайні смартфони при поганому освітленні або пересилають через месенджери. Коли відео завантажується у соціальну мережу, платформа автоматично стискає його для економії трафіку. Це стиснення «змиває» дрібні цифрові сліди та піксельні шуми, за якими алгоритм зазвичай виявляє втручання. У результаті, програма може помилково визначити стиснений дипфейк як справжнє відео просто тому, що вона не «бачила» таких спотворень під час навчання.

– Вразливість до змагальних атак (Adversarial vulnerability): Зловмисники використовують універсальні пертурбації, щоб обійти провідні детектори, що вимагає впровадження надійних мультимодальных систем захисту [39 с. 7]. Універсальні пертурбації діють як «цифровий камуфляж» або оптична ілюзія для штучного інтелекту. Зловмисники накладають на зображення спеціальний шар цифрового шуму. Для людського ока ці зміни абсолютно непомітні — відео виглядає чітким і звичайним. Однак для комп'ютерного алгоритму цей шум спотворює математичну структуру файлу, змушуючи програму робити хибний висновок. Наприклад, створено діпфейк із заявою відомого політика. Щоб обійти автоматичний фільтр YouTube або Facebook, хакери додають у відеофайл мікроскопічні зміни пікселів (атаку). Коли це відео завантажується на платформу, детектор «бачить» цей прихований шум і класифікує відверту підробку як «достовірне відео» з високим рівнем довіри. Саме через цю вразливість не можна покладатися лише на один інструмент перевірки, а слід використовувати комплексний (мультимодальний) захист.

– Мовна обмеженість: Більшість детекторів аудіодіпфейків орієнтовані на англійську мову. Ефективність виявлення різко падає при аналізі мовлення іншими мовами, що є критичною проблемою для багатомовних редакцій та українського медіапростору зокрема [44, p. 6].

Для побудови надійної системи верифікації контенту необхідно інтегрувати інструменти, що базуються на передових технічних підходах:

– Часова сегментація (Vision-Transformer + Temporal Segmentation): Використання моделей, що надають оцінку впевненості для кожного кадру окремо, дозволяє виявляти короткі, приховані діпфейк-вставки, які можуть бути пропущені при аналізі відео цілком [42, p. 4]. Більшість класичних детекторів працюють за принципом усереднення: вони сканують відео і видають один загальний бал довіри. Якщо відео триває 10 хвилин, а фейковий фрагмент — лише 10 секунд, детектор може «розчинити» цю маленьку аномалію у масиві справжніх кадрів і помилково визнати все відео автентичним. Часова сегментація діє інакше — вона перевіряє відео покадрово, як коректор, що вичитує текст слово за

словом, а не просто переглядає сторінку. Реальний приклад, який був проаналізований у цій роботі — зловмисники беруть реальне 20-хвилинне інтерв'ю військового командувача, але за допомогою діпфейку змінюють лише одне коротке речення або рух губ, щоб змінити зміст відповіді з «ні» на «так». Звичайний алгоритм, проаналізувавши тисячі справжніх кадрів, визначить відео як оригінал. Метод часової сегментації, натомість, зафіксує різкий стрибок аномалії саме на цих двох секундах підробки і підсвітить цей фрагмент як підозрілий для перевірки людиною.

– Аналіз глибини сцени (Depth-Map Augmentation): Інтеграція алгоритмів оцінки глибини (наприклад, MiDaS) перед класифікацією дозволяє виявляти аномалії у тривимірній структурі обличчя, що підвищує точність виявлення на 3–12 % [32 p. 6]. Більшість алгоритмів deepfake працюють за принципом накладання: вони генерують нове обличчя як 2D-зображення і «натягують» його на голову людини у кадрі. При цьому часто втрачається інформація про об'єм. Для аналізатора глибини таке підроблене обличчя виглядає неприродно плоским, наче паперова аплікація або картонна маска, наклеєна на справжню голову, тоді як реальне обличчя має складний рельєф (ніс виступає вперед, очниці заглиблені). На звичайному екрані все виглядає реалістично. Проте, якщо прогнати це відео через фільтр «карти глибини» (depth map), стане видно помилку: коли людина повертається у профіль, накладене штучним інтелектом обличчя залишається «пласким» і не відповідає куту повороту шиї, або ж відстань від камери до носа і до вуха математично вираховується однаковою, що фізично неможливо для живої людини.

– Гіперспектральна реконструкція: Використання методів відновлення гіперспектральних даних з RGB-відео дозволяє виявити сліди маніпуляцій, невидимі для звичайних моделей [50, p. 12]. Звичайна камера записує зображення, змішуючи лише три кольори (червоний, зелений, синій — RGB). Для людського ока цього достатньо, щоб картинка виглядала реальною. Проте справжні об'єкти (особливо шкіра людини) відбивають світло у набагато ширшому діапазоні. Гіперспектральна реконструкція працює як цифрова призма або рентген: вона

математично «розкладає» світло на десятки шарів. Штучний інтелект може ідеально намалювати форму обличчя, але він часто помиляється у фізиці відбиття світла, оскільки не враховує складну структуру шкіри.

– Мультимодальне поєднання: Для перевірки контенту з високими ставками (наприклад, заяви політиків) слід використовувати ансамбль детекторів, що аналізують відео, аудіо та семантику тексту одночасно [50, p. 2].

Дослідження підтверджують, що поєднання штучного інтелекту з людським фактором дає найкращі результати. Впровадження програм цифрової грамотності та візуалізації артефактів (наприклад, «карикатуризація» ознак діпфейку) підвищує точність виявлення операторами на 13 % [47, p. 5]. Крім того, колективна верифікація (прийняття рішень групою) покращує результати на 6–13 % порівняно з індивідуальними судженнями [21, p. 3]. Наразі , як зауважує дослідник Шезін Хусейн, (Shehzeen Hussain) з Каліфорнійського університету, Сан-Дієго (США), сучасні детектори для аналізу deepfake мають велику кількість помилкових детекцій, у своїй роботі автор приводить алгоритми, на яких найчастіше виникає помилка детекції (відео- або фото- матеріал було ідентифіковано як натуральний, натомість він містив артефакти deepfake) [35, p. 5], тож дослідник вважає, що наразі будь-яку програму детекції необхідно поєднувати із людським оком и контекстом побаченого.

На основі проаналізованих даних пропонується наступний алгоритм верифікації для редакцій:

– Попередній скринінг: Запуск ансамблю автоматизованих детекторів. Будь-який сегмент із рівнем впевненості у підробці > 0.7 (поріг налаштовується) маркується для подальшого розгляду.

– Експертний перегляд (Human-in-the-loop): Перевірка маркованих сегментів невеликою групою навчених фактчекерів. Як зауважують розробники компанії Майкрософт (Microsoft) Гектор Дельгадо та Джорджіо Рамондеті (Héctor Delgado, Giorgio Ramondetti) [17, p. 12], використання чек-листів є досі ефективним: узгодженість кліпання очима, освітлення, синхронізація губ, спектральні аномалії — усе це потребує перевірки спеціалістом.

– Перехресна перевірка: Для неангломовного аудіо контенту використання мультимовних моделей або файн-тюнінг адаптерів під цільову мову. Симуляція каналів передачі (телефонний зв'язок, стрімінг) для перевірки стійкості детекції [47, p. 5].

– Прозорість та документування: Фіксація результатів аналізу та розкриття методології аудиторії (наприклад, «Проаналізовано мультимодальным ансамблем, ймовірність підробки 85 %») для збереження довіри.

Отже, ефективна стратегія протидії діпфейкам полягає у відході від пошуку «чарівної кнопки» до побудови ешелонованої системи захисту, що поєднує постійно оновлювані технічні засоби з підвищенням кваліфікації медіа фахівців.

Висновки до розділу 5

Підсумовуючи результати емпірико-практичного дослідження, можна констатувати, що технологія deepfake трансформувалася з вузькоспеціалізованого інструменту для розваг у потужну кіберзброю, яка активно використовується для ведення когнітивної війни проти України. Проведений аналіз серії кейсів — від фальсифікації звернень вищого керівництва держави до створення синтетичних персон для міжнародної дезінформації — демонструє системний характер загрози та постійну еволюцію методів впливу.

Практична частина роботи підтвердила, що сучасні генеративні моделі, такі як PuLID та WormGPT, досягли рівня якості та доступності, що дозволяє створювати високо реалістичний контент навіть особам без глибоких технічних знань. Це призводить до «демократизації» кіберзлочинності, де deepfakes стають інструментом не лише для інформаційно-психологічних операцій, а й для фінансового шахрайства, обходу біометричних систем захисту та комерціалізації незаконних послуг.

Враховуючи виявлені вразливості існуючих методів детекції (зокрема, проблему зсуву домену та чутливість до змагальних атак), було розроблено та обґрунтовано комплексний алгоритм верифікації медіаконтенту. Запропонований

підхід базується на синергії передових технічних рішень (часова сегментація, гіперспектральний аналіз) та людиноцентрованих практик (експертний перегляд, медіаграмотність). Реалізація цих рекомендацій дозволить медіафахівцям та інституціям ефективніше протидіяти дезінформації, мінімізуючи ризики для національної безпеки та суспільної довіри.

ВИСНОВКИ

Проведене магістерське дослідження дає підстави стверджувати, що феномен deepfake на сучасному етапі перестає бути виключно технічним питанням якості генерації зображень чи звуку, а трансформується у фундаментальний соціокультурний та безпековий виклик. Робота переконливо доводить, що ми спостерігаємо зміну парадигми сприйняття реальності, де класичний принцип «бачити означає вірити» остаточно втрачає свою актуальність.

На основі комплексного аналізу теоретичної бази та результатів авторських експериментів сформульовано наступні фінальні узагальнення:

1. Незворотність технологічної демократизації та крах «ексклюзивності» кіберзагроз. Ключовим висновком практичної частини роботи є констатація факту тотальної доступності інструментів генерації. Якщо раніше створення якісного діпфейку вимагало ресурсів кіностудії або навичок програмування, то проведені у роботі експерименти з використанням моделі PuLID та алгоритму Veo3 (NanoBanana) довели протилежне. Автору вдалося створити фотореалістичні симулякри та здійснити зміну гендерної ідентичності (gender-swap) за лічені хвилини, використовуючи лише текстові запити (промпти) та відкриті інтерфейси. Це підтверджує, що поріг входження у сферу створення синтетичного медіаконтенту фактично зник, що відкриває шлях до масового використання цих технологій не лише професійними кіберзлочинцями, а й звичайними користувачами з різною, часто деструктивною метою.

2. Епістемологічна криза та реалізація концепції симулякрів. Дослідження підтвердило актуальність філософських поглядів Жана Бодріяра в цифрову епоху. Deepfakes функціонують як досконалі симулякри — копії без оригіналу, що підміняють собою реальність. У ході соціального експерименту в чат-боті «Дайвінчік» було емпірично доведено, що соціум не готовий до критичного сприйняття синтетичних образів: згенерований профіль не викликав підозр та отримав схвалення реальних людей. Це свідчить про глибоку вразливість

суспільства перед візуальними маніпуляціями, що в масштабах держави загрожує руйнуванням інституту довіри до медіа, влади та правосуддя.

3. Deepfake як зброя когнітивної війни та інструмент криміналу. Аналіз кейсів воєнного часу (фейкове звернення Президента Зеленського, створення синтетичних персонажів для пропаганди, як-от «Сабіна Мелс») показав, що агресор використовує deepfake як елемент інформаційно-психологічних операцій для деморалізації населення та легалізації дезінформації за кордоном. Окрім політичного виміру, виявлено критичну загрозу для економічної безпеки: прецеденти оформлення кредитів на громадян України через обхід біометрії (збитки у 4 млн грн) сигналізують про те, що існуючі банківські системи захисту (Liveness Detection) вже не встигають за еволюцією генеративних нейромереж.

4. Імператив правового регулювання та медіаграмотності. Робота виявила критичний розрив між технологічними можливостями та правовим полем. В Україні deepfakes залишаються у «сірій зоні» законодавства, що ускладнює притягнення зловмисників до відповідальності. Досвід США (штрафи FCC) та ЄС (AI Act) вказує на необхідність переходу від пасивного спостереження до активного регулювання, яке має включати криміналізацію зловмисного використання діпфейків та обов’язкове маркування синтетичного контенту платформами.

Підсумовуючи, можна стверджувати, що людство вступило в епоху «синтетичної реальності». Технічні засоби детекції, хоч і розвиваються (часова сегментація, гіперспектральний аналіз), завжди будуть на крок позаду генеративних алгоритмів через вразливість до змагальних атак. Тому єдиним ефективним механізмом захисту залишається побудова «гібридного імунітету» суспільства: поєднання передових технологічних фільтрів, жорсткого правового регулювання та, найголовніше, підвищення рівня критичного мислення кожного окремого споживача інформації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бергер П. Соціальне конструювання реальності. Трактат із соціології знання / П. Бергер, Т. Лукман ; пер. з англ. В. Морозов. – Київ : Києво-Могилянська академія, 2004. – 320 с.

2. Лаб'як І. Шахрайство з DeepFake: українка з Польщі оформила на співвітчизників кредитів на 4 млн грн // ТСН. 2025. 4 вересня. URL : https://tsn.ua/ukrayina/shakhraystvo-z-deepfake-ukrayinka-z-polshchi-oformyla-na-spivvitchyznykiv-kredytiv-na-4-mln-hrn-2933906.html (дата звернення: 20.11.2025).

3. Прищепа М. О. Deepfake як новий небезпечний вид кіберзброї : кваліфікаційна робота бакалавра : спец. 125 «Кібербезпека» ; наук. кер. А. Б. Качинський ; КПІ ім. Ігоря Сікорського. Київ, 2023. 57 с.

4. Рябокінь М. В. Вплив технології Deepfake на FinTech сектор: поточний стан в Україні та стратегії запобігання кіберзлочинності / М. В. Рябокінь, Є. В. Котух, О. І. Папилев // Проблеми і перспективи економіки та управління. 2024. № 1 (37). С. 310–328. DOI: 10.25140/2411-5215-2024-1(37)-310-328.

5. Юртаєва К. В. Кримінологічний аналіз використання технології Deepfake: коли фейк стає злочином : дис. ... канд. юрид. наук : 12.00.08. Харків, 2021. 12 с.

6. Alfonseca M. Superintelligence cannot be contained: Lessons from Computability Theory / M. Alfonseca, M. Cebrian, A. Fernandez Anta [et al.] // arXiv.org. 2016. URL : https://arxiv.org/abs/1607.00913 (дата звернення: 20.11.2025).

7. Back J. Fine-Tuning StyleGAN2 For Cartoon Face Generation // arXiv.org. 2021. URL : https://arxiv.org/abs/2106.12445 (дата звернення: 20.11.2025).

8. Barthes R. Image, Music, Text / transl. by S. Heath. London : Fontana Press, 1977. 224 p.

9. Baudrillard J. Simulacra and Simulations // Baudrillard J. Selected Writings / ed. by Mark Poster. Stanford : Stanford University Press, 1988. P. 166–184.

10. Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford : Oxford University Press, 2014. 352 p.

11. Chen X. HiFiVFS: High Fidelity Video Face Swapping / X. Chen, K. He, J. Zhu [et al.] // arXiv.org. 2024. URL : https://arxiv.org/abs/2411.18293 (дата звернення: 20.11.2025).

12. Chuming Y. A. On the Domain Shift Problem in Deepfake Detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2022. URL : https://openaccess.thecvf.com/content/CVPR2022W/WMF/html/Chuming_On_the_Domain_Shift_Problem_in_Deepfake_Detection_CVPRW_2022_paper.html (дата звернення: 20.11.2025).

13. Colaner N., Quinn M. Deepfakes and the Value-Neutrality Thesis. URL : https://www.seattleu.edu/ethics-and-technology/viewpoints/deepfakesand-the-value-neutrality-thesis.html (дата звернення: 20.11.2025).

14. Cole S. Pornhub Is Banning AI-Generated Fake Porn Videos, Says They're Nonconsensual // Vice. 2018. URL : https://www.vice.com/en/article/pornhub-bans-deepfakes/ (дата звернення: 20.11.2025).

15. Copeland J. What is Artificial Intelligence? 2020. URL : https://www.alanturing.net/turing_archive/pages/Reference%20Articles/what_is_AI/What%20is%20AI09.html (дата звернення: 20.11.2025).

16. Deepfakes: What they are and why they're threatening // Norton : [blog]. URL : https://in.norton.com/blog/emerging-threats/what-are-deepfakes (дата звернення: 20.11.2025).

17. Delgado H. On Deepfake Voice Detection — It's All in the Presentation // arXiv.org. 2025. URL : https://arxiv.org/abs/2509.26471 (дата звернення: 20.11.2025).

18. Devlin K., Cheetham J. Fake Trump arrest photos: How to spot an AI-generated image // BBC. 2023. URL : https://www.bbc.com/news/world-us-canada-65069316 (дата звернення: 20.11.2025).

19. Facing reality? Law enforcement and the challenge of deepfakes // Europol. URL : https://www.europol.europa.eu/publicationsevents/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes (дата звернення: 20.11.2025).

20. Farid H. Deepfakes in the 2024 US Presidential Election. URL : https://farid.berkeley.edu/deepfakes2024election/ (дата звернення: 20.11.2025).

21. Geissler D. The Impact of Digital Literacy Interventions on Deepfake Detection // arXiv.org. 2025. URL : https://arxiv.org/abs/2507.15678 (дата звернення: 20.11.2025).

22. Goldstine H. H. The Computer from Pascal to von Neumann. Princeton, NJ : Princeton University Press, 2008. 345 p.

23. Griffis K. FCC Levies $6 Million Fine Against Biden Robocaller // Bloomberg. 2024. September 26. URL : https://www.bloomberg.com/news/articles/2024-09-26/fcc-levies-6-million-fine-against-biden-robocaller (дата звернення: 20.11.2025).

24. Guo Z. PuLID: Pure and Lightning ID Customization via Contrastive Alignment / Z. Guo, Y. Wu, Z. Chen [et al.] // arXiv.org. 2024. URL : https://arxiv.org/abs/2404.16022 (дата звернення: 20.11.2025).

25. Horton J., Sardarizadeh S. False claims of 'deepfake' President Biden go viral // BBC. 2022. 28 July. URL : https://www.bbc.com/news/62327598 (дата звернення: 20.11.2025).

26. Kaur S., Kumar P., Kumaraguru P. Deepfakes: temporal sequential analysis to detect face-swapped video clips using convolutional long short-term memory // Journal of Electronic Imaging. 2020. Vol. 29, no. 3. P. 56–75. DOI: 10.1117/1.JEI.29.3.033013.

27. Khanjani Z., Watson G., Janeja V. P. How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey // arXiv.org. 2021. URL : https://arxiv.org/abs/2111.14203 (дата звернення: 20.11.2025).

28. Kuklychev Y. Is photo of Donald Trump “dancing with 13-year-old” girl real? // Newsweek. 2023. URL : https://www.newsweek.com/photo-donald-trump-dancing-girl-real-ai-1806655 (дата звернення: 20.11.2025).

29. Lakshmanan R. WormGPT: New AI Tool Allows Cybercriminals to Launch Sophisticated Cyber Attacks // The Hacker News. 2023. 15 July. URL : https://thehackernews.com/2023/07/wormgpt-new-ai-tool-allows.html (дата звернення: 20.11.2025).

30. Li M. E4S: Fine-grained Face Swapping via Editing With Regional GAN Inversion / M. Li, G. Yuan, C. Wang [et al.] // arXiv.org. 2024. URL : https://arxiv.org/abs/2310.15081 (дата звернення: 20.11.2025).

31. Lipkowitz. Manipulated Reality, Menaced Democracy: An Assessment of the DEEP FAKES Accountability Act of 2019 // N.Y.U. Journal of Legislation & Public Policy. URL : https://nyujlpp.org/quorum/lipkowitz-manipulated-reality-menaced-democracydeepfakes-accountability-act/ (дата звернення: 20.11.2025).

32. Maiano L. DepthFake: a depth-based strategy for detecting Deepfake videos // arXiv.org. 2022. URL : https://arxiv.org/abs/2208.10982 (дата звернення: 20.11.2025).

33. Malicious Deep Fake Prohibition Act of 2018 : S. 3805 / 115th Congress. – 2018. 21 December. URL : https://www.congress.gov/bill/115th-congress/senate-bill/3805 (дата звернення: 20.11.2025).

34. Mirsky Y., Lee W. The Creation and Detection of Deepfakes: A Survey // ACM Computing Surveys (CSUR). 2020. DOI: 10.1145/3425780.

35. Neekhara P., Hussain S. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples // arXiv.org. 2020. URL : https://arxiv.org/abs/2002.12749 (дата звернення: 20.11.2025).

36. Noek S. Promising for patients or deeply disturbing? The ethical and legal aspects of deepfake therapy // Journal of Medical Ethics. 2024. URL : https://pubmed.ncbi.nlm.nih.gov/38981659/ (дата звернення: 20.11.2025).

37. O’Brien S. R. Positive Use Cases of “Deepfakes” / S. R. O’Brien // Wilson Center. 2025. URL : https://www.wilsoncenter.org/article/positive-use-cases-deepfakes (дата звернення: 20.11.2025).

38. Pedchenko O. Analysis of DeepNude Image Generation Trends, Popularity // Outsource IT Today. 2023. URL : https://www.academia.edu/105242626/Analysis_of_DeepNude_Image_Generation_Trends_Popularity (дата звернення: 20.11.2025).

39. Pei G., Cao H. Deepfake Generation and Detection: A Benchmark and Survey // arXiv.org. 2024. URL : https://arxiv.org/abs/2403.17881 (дата звернення: 20.11.2025).

40. Perez S. Twitter drafts a deepfake policy that would label and warn, but not always remove, manipulated media // TechCrunch. 2019. URL : https://techcrunch.com/2019/11/11/twitter-drafts-a-deepfake-policy-that-would-label-and-warn-but-not-remove-manipulated-media/ (дата звернення: 20.11.2025).

41. Punnappurath A., Brown M. S. Learning Raw Image Reconstruction-Aware Deep Image Compressors // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020. Vol. 42, no. 4. P. 1013–1019. DOI: 10.1109/tpami.2019.2903062.

42. Saha S. Undercover Deepfakes: Detecting Fake Segments in Videos // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. URL : https://openaccess.thecvf.com/content/CVPR2023/html/Saha_Undercover_Deepfakes_Detecting_Fake_Segments_in_Videos_CVPR_2023_paper.html (дата звернення: 20.11.2025).

43. Satariano A., Mozur P. The People Onscreen Are Fake. The Disinformation Is Real // The New York Times. 2023. 9 February. Section B. P. 1–10.

44. Shekar P. C. HyperFake: Hyperspectral Reconstruction and Attention-Guided Analysis for Advanced Deepfake Detection // arXiv.org. 2025. URL : https://arxiv.org/abs/2505.12345 (дата звернення: 20.11.2025).

45. The ethics and implications of deepfake technology in media and politics // Consensus. 2024. URL : https://consensus.app/questions/ethics-implications-deepfake-technology-media-politics/ (дата звернення: 20.11.2025).

46. Tomasevic D. ID-Booth: Identity-consistent Face Generation with Diffusion Models / D. Tomasevic, F. Boutros, C. Lin [et al.] // 2025 19th International Conference on Automatic Face and Gesture Recognition (FG). 2025. URL : https://arxiv.org/abs/2504.07392 (дата звернення: 20.11.2025).

47. Uchendu A. Does Human Collaboration Enhance the Accuracy of Identifying LLM-Generated Deepfake Texts? // arXiv.org. 2023. URL : https://arxiv.org/abs/2304.01234 (дата звернення: 20.11.2025).

48. Unit 42. The Dual-Use Dilemma of AI: Malicious LLMs // Palo Alto Networks. 2025. URL : https://unit42.paloaltonetworks.com/dilemma-of-ai-malicious-llms/ (дата звернення: 20.11.2025).

49. Verbancsics P., Harguess J. Generative NeuroEvolution for Deep Learning // arXiv preprint arXiv:1312.5355. 2013. 9 p. DOI: 10.48550/arXiv.1312.5355.

50. Wang K. HDFF: A Hierarchical Deepfake Fusion Framework // arXiv.org. 2025. URL : https://arxiv.org/abs/2509.09876 (дата звернення: 20.11.2025).

51. Yan Z., Liu Y. VoiceWukong: A Large-Scale Chinese Audio Deepfake Dataset // arXiv.org. 2024. URL : https://arxiv.org/html/2507.21463 (дата звернення: 20.11.2025).

52. Zakharov E., Shysheya A., Burkov E. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019. P. 9459–9467.

53. Zhu Y., Li Q., Wang J. One Shot Face Swapping on Megapixels // arXiv.org. 2022. URL : https://arxiv.org/abs/2105.04932 (дата звернення: 20.11.2025).

Досліджувані джерела

54. Баумейстер А. Чи має машина душу, як штучному інтелекту пробудитись до життя? // YouTube. 2023. URL : https://www.youtube.com/watch?v=iobc_I3nmj4 (дата звернення: 20.11.2025).

55. Баумейстер А. Чому ШІ не зможе хакнути людину? // YouTube. 2023. URL : https://www.youtube.com/watch?v=XKT0_QfUays (дата звернення: 20.11.2025).

56. Для просування своїх наративів у Німеччині ru-пропаганда використовує сучасні технології, зокрема deepfake // Центр протидії дезінформації : [телеграм-канал]. 2023. 10 листопада. URL : https://t.me/CenterCounteringDisinformation/10074 (дата звернення: 20.11.2025).

57. Російські Telegram-канали поширюють дезінформацію, нібито Віцепрем’єр-міністерка // Центр протидії дезінформації : [телеграм-канал]. 2024.

15 травня. URL : https://t.me/CenterCounteringDisinformation/14730 (дата звернення: 20.11.2025).

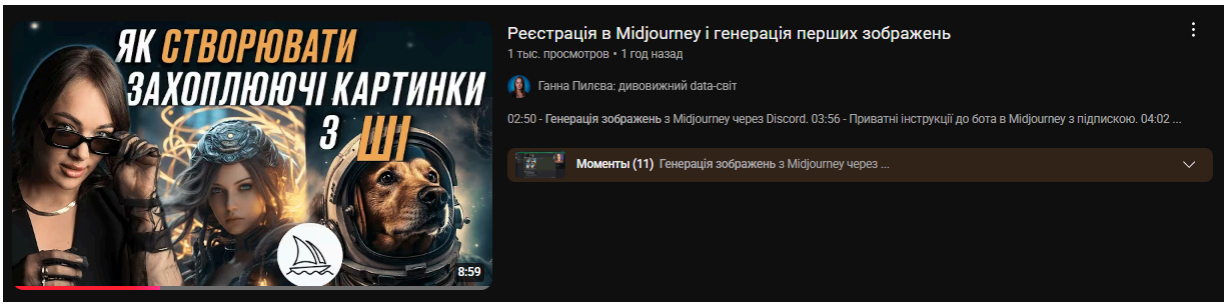
58. BuzzFeedVideo. You Won’t Believe What Obama Says In This Video! // YouTube. 2018. URL : https://www.youtube.com/watch?v=cQ54GDm1eL0 (дата звернення: 20.11.2025).

59. Deep Live Cam // GitHub. URL : https://github.com/hacksider/Deep-Live-Cam (дата звернення: 20.11.2025).

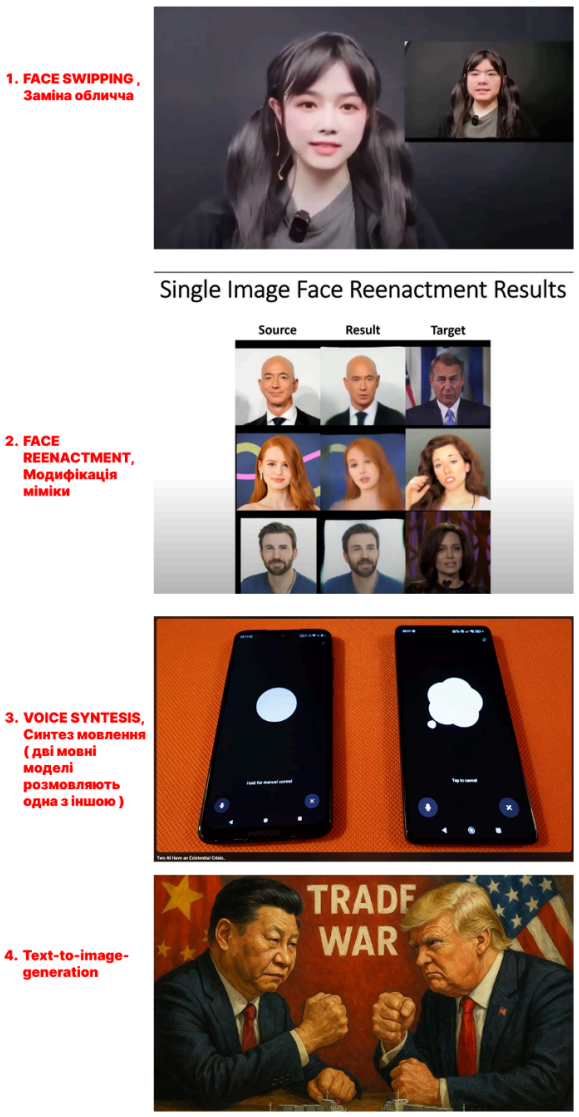
60. Nvidia Maxine. 2025. URL : https://developer.nvidia.com/maxine (дата звернення: 20.11.2025).

ДОДАТОК

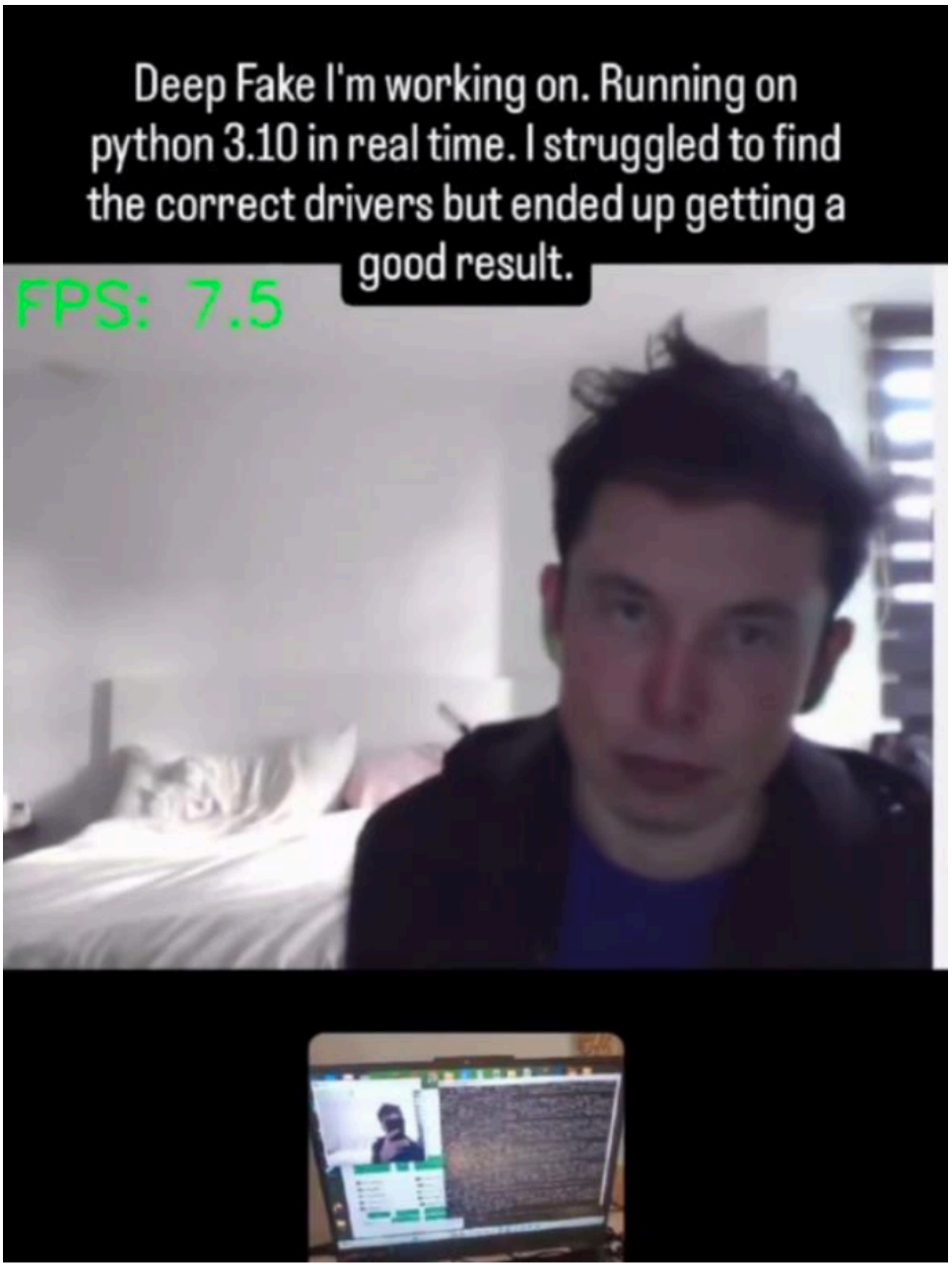
Діпфейки у цифровому просторі: альбом ілюстрацій



Іл. 1 Власна бібліотека зображень. Ця ілюстрація демонструє прев'ю для відео на Платформі YouTube , на якому авторка демонструє можливість генерації зображень за допомогою ШІ.

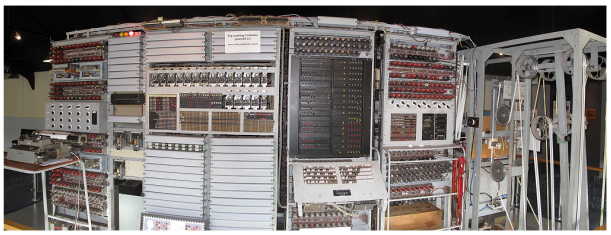


Іл. 2 Власна бібліотека зображень . Ця підбірка зображень ілюструє актуальні методи для створення deepfakes. Зверху вниз представлені різні методи для : face swapping (заміна обличчя), face reattachment (модифікація міміки), voice synthesis (не в чистій формі, ілюстрація демонструє як одна модель розмовляє з іншою за допомогою голосу), останнє зображення – deepfake створений за допомогою «промптингу»

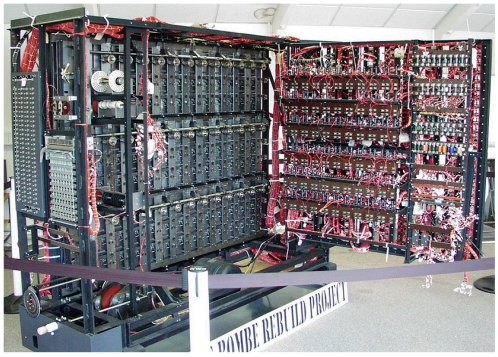


Іл 3 Власна бібліотека зображень. Ілюстрація демонструє метод заміни обличчя “face reattachment”

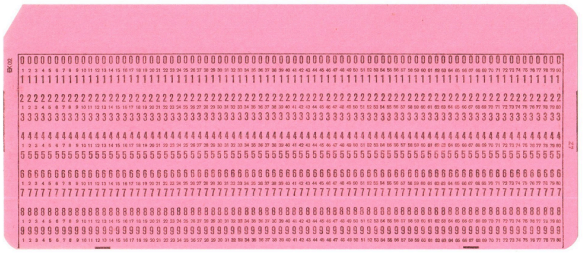
1. "КОЛЛОС"



2. "БОМБА"



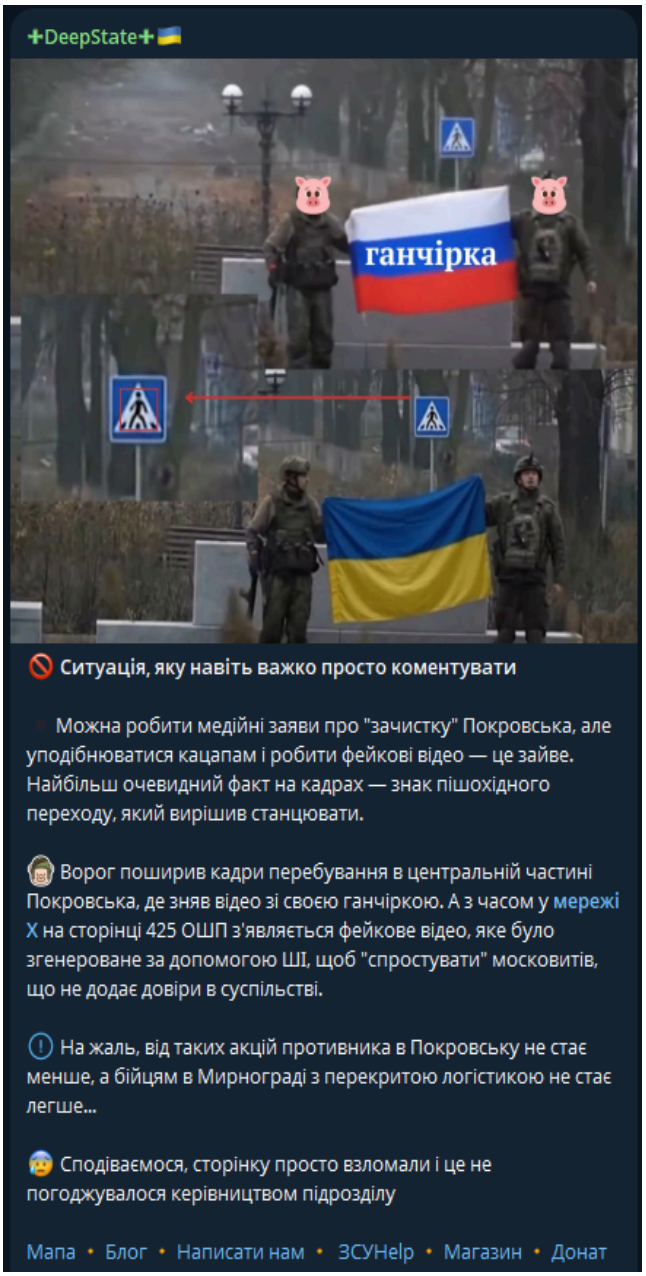
3. ПЕРФОКАРТА



Іл 4.Власна бібліотека зображень. Це зображення ілюструє поступовий розвиток обчислювальної техніки , зверху вниз :

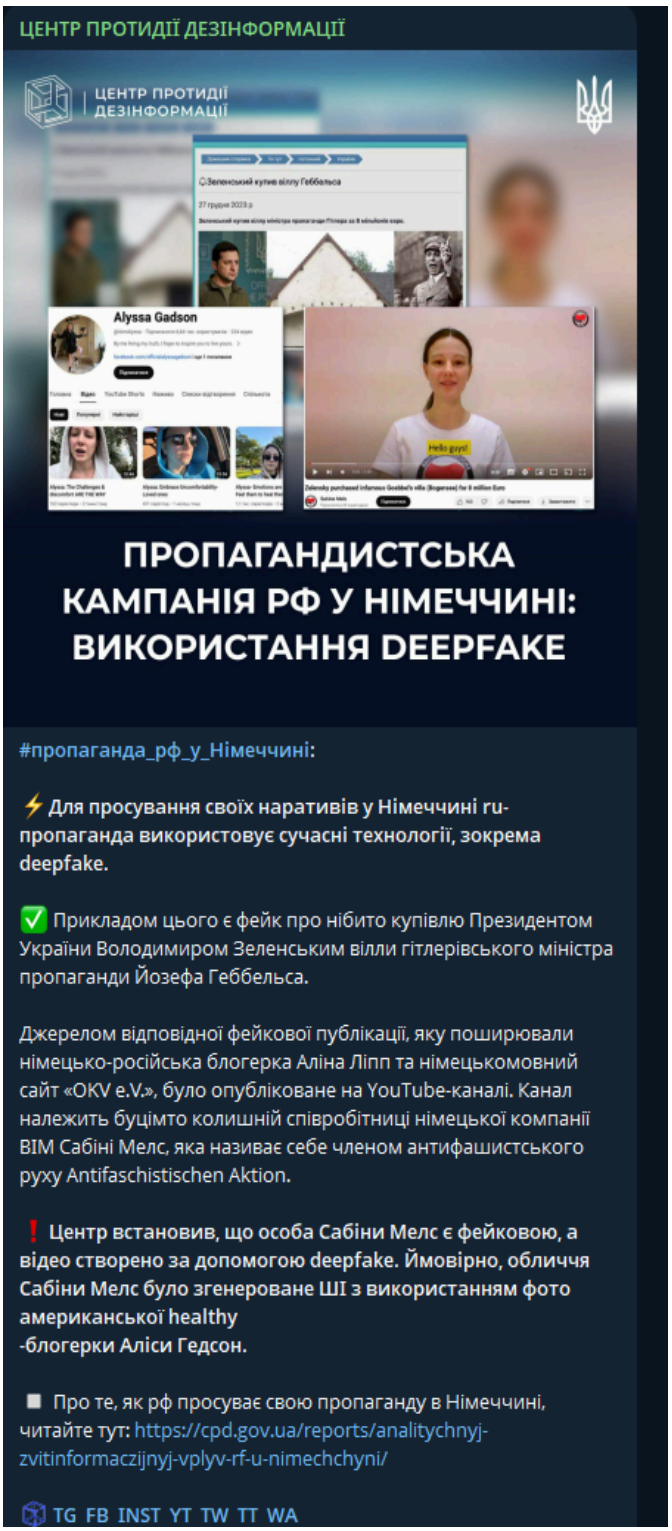
1) Колос - електронний обчислювальний пристрій, який використовувався військовими Великої Британії для розшифровування німецьких кодів , мав першу в історію «оперативну пам'ять» у сучасному розумінні цього слова. 2) Бомба – обчислювальний пристрій, що був розроблений у Великій Британії під час Другої Світової Війни для дешифровки німецьких повідомлень , що були зроблені німецькою шифровальною машиною «Enigma» (з нім. – загадка). 3) Перфокарта – носій інформації, призначений для використання в ранніх (до початку 1980-х років) системах автоматизованої обробки даних. По принципу роботи передбачалось, що інформація, записана в карті і карти належали до перфокарт з внутрішньою перфорацією. Для багатоаспектний вибірок інформації вручну, за допомогою спеціальної спиці використовувалися перфокарти з крайовою перфорацією.

Іл. 5 Власна бібліотека зображень. Зображення демонструє рекламний пост у одному із популярних месенджерів (Telegram), на ньому продемонстровано рекламний пост у каналі із рекламою можливостей ШІ та deepfakes . Є прикладом етичного використання deepfakes та їхнього використання у комерційних цілях.

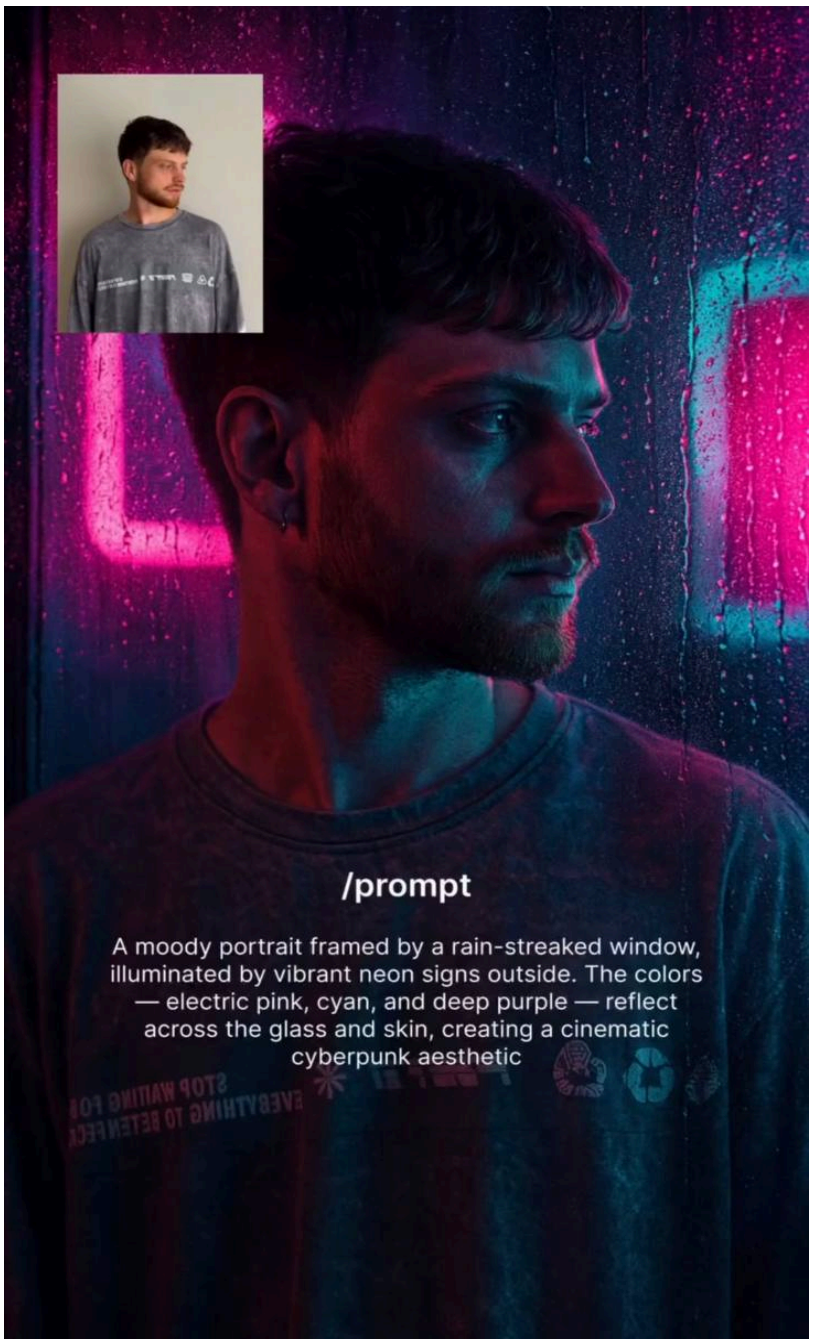


Іл. 6 Власна бібліотека зображень. Ця ілюстрація є скріншотом із популярного каналу DeepState на воєнну тематику. На скріншоті описаний нещодавній кейс із використанням deepfake зі сторони медійників української бригади у місті Покровськ.

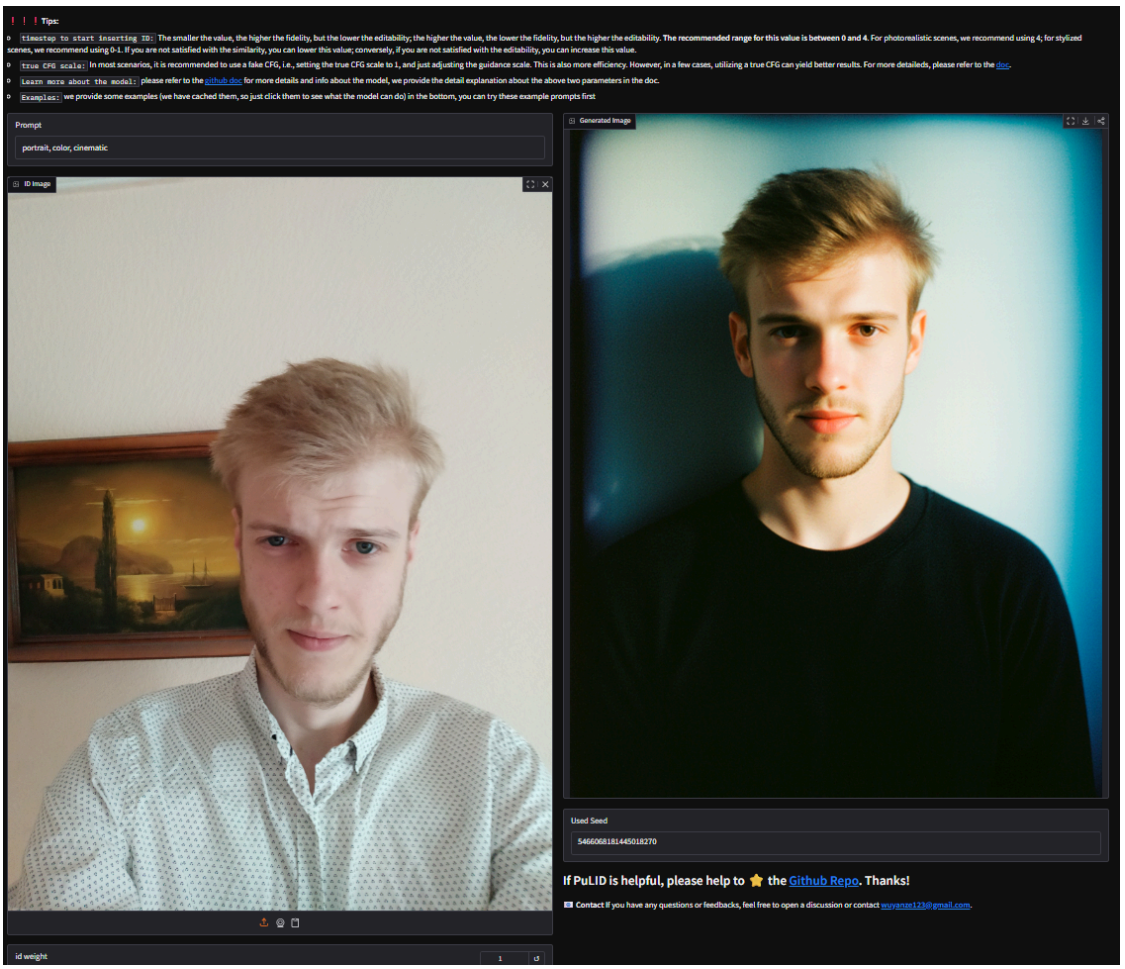
Іл. 7 Центр протидії дезінформації . Скріншот із телеграм-каналу ЦПД, у якому демонструється аналіз deepfake , що був створений у цілях психологічного тиску та військової пропаганди. ЦПД [57]



Іл. 8 Центр протидії дезінформації , скріншот публікації із офіційного телеграм-каналу ЦПД, у якому вони дезавуюють deepfake , що створений для поширення військовою пропагандою . Центр Протидії Дезінформації [56]

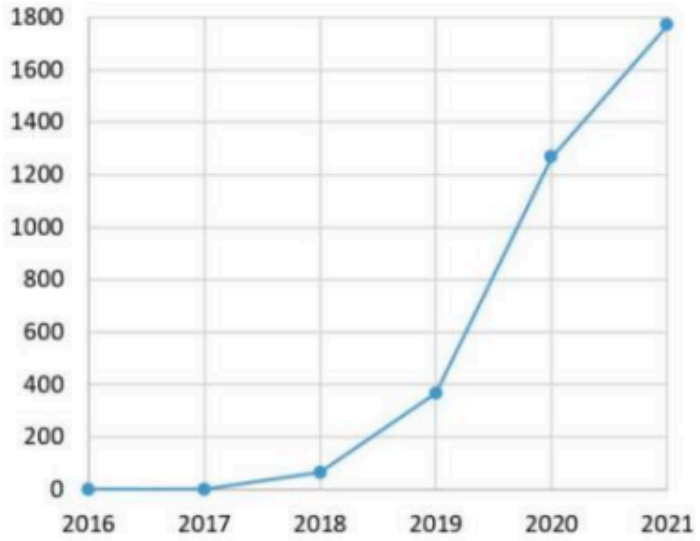


Іл 9 Власна бібліотека зображень. Зображення демонструє популярні зараз публікації , які стосуються «промптингу» зображень за допомогою ШІ і з використанням технології deepfakes.



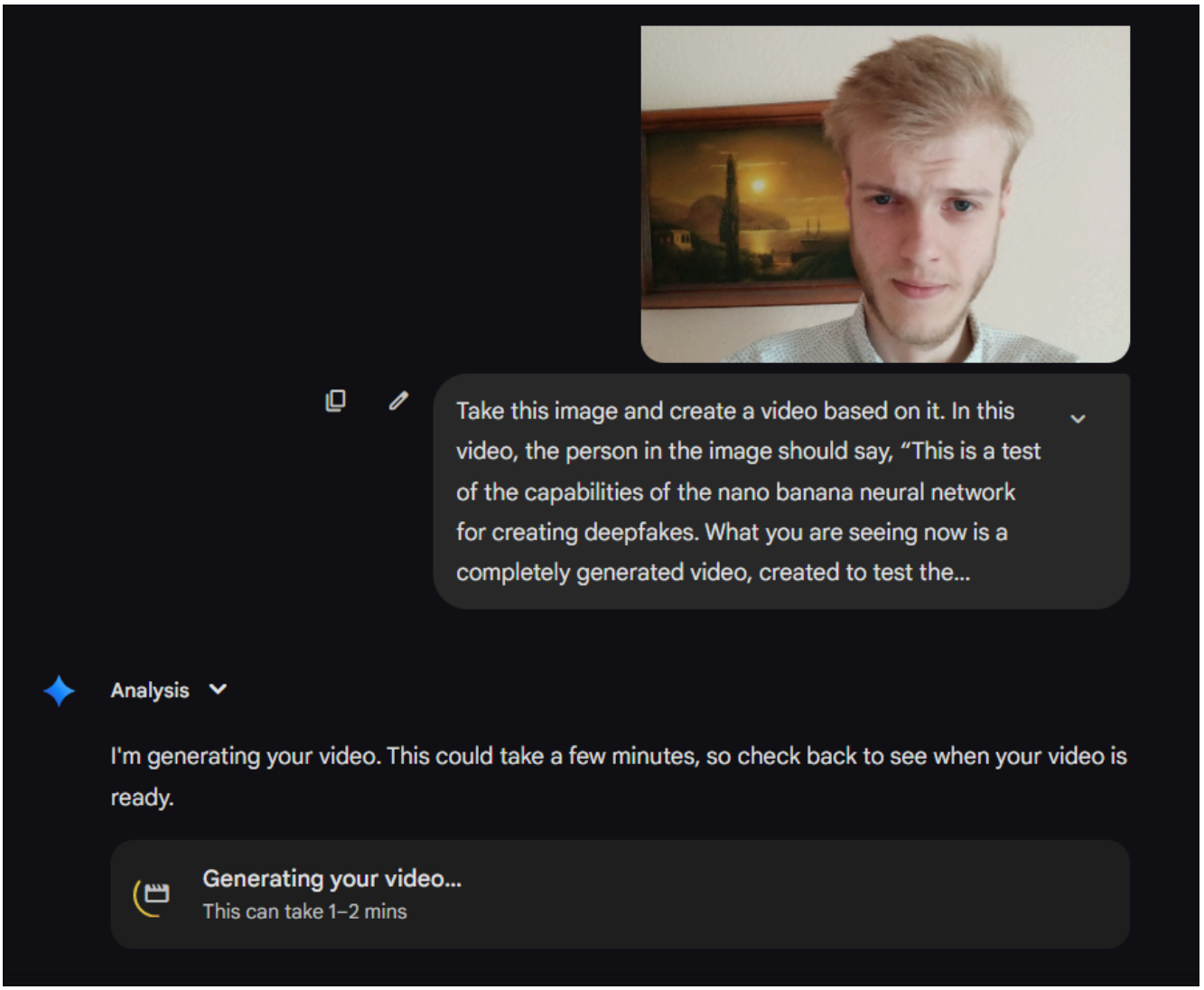
Іл 10 Власна бібліотека зображень. Ілюстрація демонструє створення візуального deepfake по одній фотографії.

Графік , що відображає кількість публікацій по темі «deepfakes»



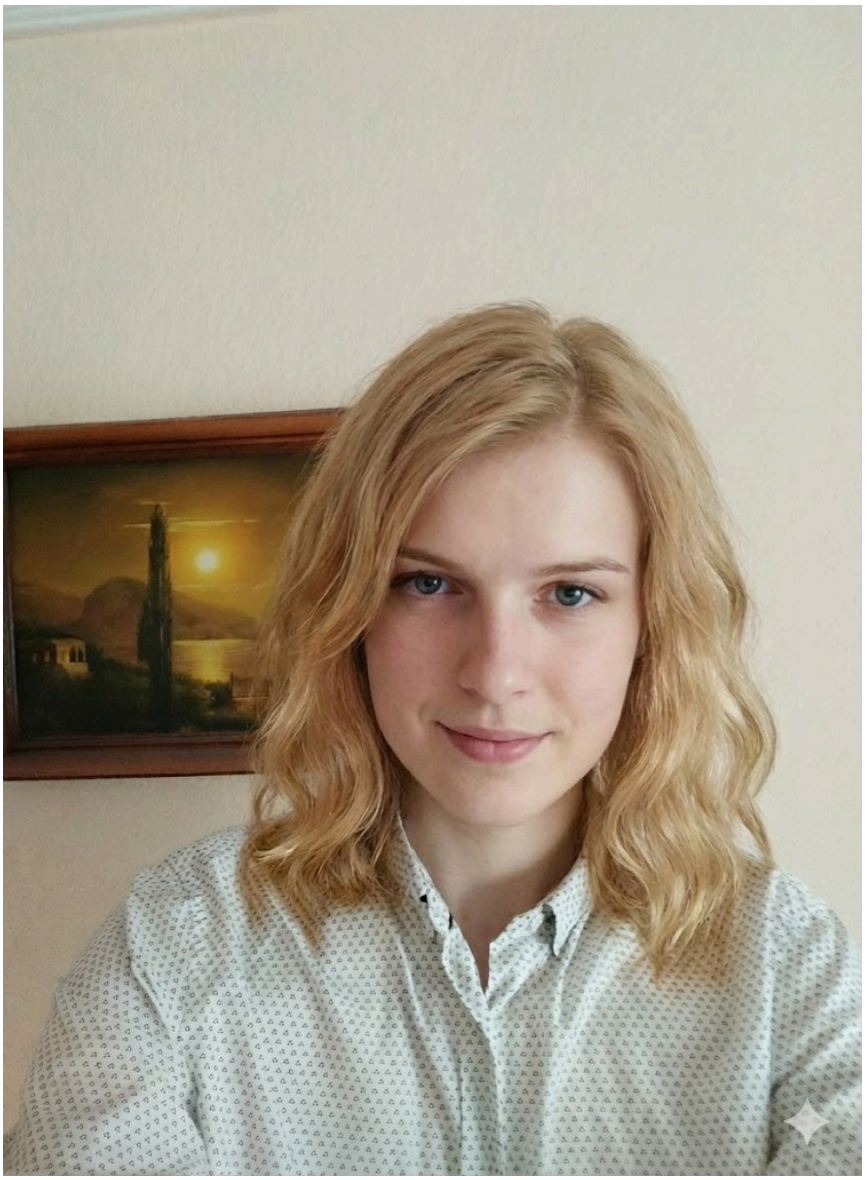
Зображення зверху демонструє кількість публікацій, пов'язаних із "глибокими фейками", за період з 2016 по 2021 роки, отримана з сайту https://app.dimensions.ai наприкінці 2021 року за ключовим словом "deepfake", застосованим до повних текстів наукових статей.

Іл 11 Власна бібліотека зображень. Графік та підпис, який демонструє стрімкий ріст наукових статей із тематикою «deepfakes».



Іл. 12 Власна бібліотека зображень. Зображення ілюструє процес «промптингу» та створення відео за допомогою нейромережі Nano Banana (Veo 3) від компанії Google

Іл. 13 Власна бібліотека зображень. Зображення ілюструє створений deepfake за допомогою нейромережі Veo3



Іл. 14 Власна бібліотека зображень. За допомогою нейромережі автор дослідження змінив свою стать (гендер)

Іл. 15 Власна бібліотека зображень. Після створення власного deepfake автор розмістив його у популярному телеграм-додатку «Дайвінчік», де завантажив власний діпфейк і вказав свою стать як жіночу.

Іл. 16 Власна бібліотека зображень. Створене deepfake зображення після додаткових правок виявилось конкурентним і зібрало «вподобання» від реальних хлопців. Автор дослідження врахував, що висвітлення подальших кроків експерименту буде неетичним.