СИНТЕТИЧНІ ЦИФРОВІ ДОКАЗИ В УМОВАХ ВІЙНИ: ТЕХНОЛОГІЇ ВИЯВЛЕННЯ, ВЕРИФІКАЦІЇ ТА ДОКУМЕНТУВАННЯ AI-ЗГЕНЕРОВАНОГО КОНТЕНТУ

Ольга ГАБОРЕЦЬ
доктор філософії, доцент,
доцент кафедри оперативно-розшукової діяльності
та інформаційної безпеки,
Донецький державний університет внутрішніх справ

Поширення генеративного штучного інтелекту змінює не лише структуру інформаційних загроз, а й самі умови встановлення факту в кримінальному провадженні. Фото-, аудіо- та відеоматеріал більше не може презюмуватися достовірним лише тому, що він правдоподібно відтворює зовнішність, голос, місце або перебіг події. В умовах війни синтетичний контент здатний імітувати наказ військового командування, звернення посадової особи, наслідки бойового ураження, присутність людини у певній локації чи її комунікацію з представниками держави-агресора. Водночас діпфейки в умовах інформаційної війни становлять загрозу національній безпеці [2] і створюють протилежний ризик: справжній запис може бути оголошений «згенерованим», щоб дискредитувати доказ та уникнути відповідальності. Отже, правоохоронна система стикається не з однією, а з двома симетричними загрозами – прийняттям синтетичного матеріалу за автентичний та відхиленням автентичного матеріалу як синтетичного.

Саме цей подвійний ризик становить нетривіальний вимір проблеми. Б. Чесні та Д. Сітрон визначили можливість заперечувати справжні записи через загальну обізнаність про deepfake як «дивіденд брехуна» [3]. Експеримент К. Ваккарі й Е. Чедвіка показав, що політичні deepfake частіше породжують невизначеність, ніж безпосередньо вводять аудиторію в оману, однак така невизначеність знижує довіру до новин у соціальних мережах [4]. Дослідження Дж. Тумі та співавторів, проведене на матеріалі обговорень deepfake під час російсько-української війни, засвідчило послаблення довіри

користувачів до можливості встановити достовірність воєнних відеоматеріалів, що автори концептуалізують як підрив епістемічної довіри [5]. Таким чином, синтетичний контент є не лише підробленим повідомленням, а технологією руйнування критеріїв, за якими суспільство та інституції відрізняють документ від симуляції.

Метою дослідження є обґрунтування комплексного підходу до виявлення, верифікації та документування AI-згенерованого контенту, що потенційно має доказове значення, а також розроблення авторської п'ятикомпонентної моделі оцінювання його походження й автентичності в умовах війни. Методологічну основу становлять системний і ризик-орієнтований підходи, методи цифрової криміналістики, OSINT, порівняльного аналізу, логічної тріангуляції та оцінювання надійності джерел.

Термін «синтетичний цифровий доказ» потребує методологічного застереження. Штучно створений файл не набуває статусу доказу автоматично: до процесуального оцінювання коректніше говорити про синтетичний цифровий матеріал із потенційним доказовим значенням. При цьому синтетичне походження не завжди виключає доказову цінність. Deepfake може бути предметом протиправного впливу, знаряддям шахрайства, способом погрози або елементом інформаційної операції; тоді доказове значення має не правдивість зображеної події, а факт створення, поширення, адресування та впливу відповідного об'єкта. Тому верифікація повинна відповідати на два різні питання: чи відтворює матеріал реальну подію та що доводить саме існування цього цифрового об'єкта.

Технічна складність полягає у відсутності стабільного універсального маркера синтетичного походження. І. Мірскі та В. Лі систематизували методи створення й виявлення deepfake та показали їх залежність від конкретної архітектури генерації [6]. Л. Вердоліва наголошує, що медіафорензика працює з різними групами слідів: фізичними, сенсорними, статистичними, семантичними та слідами монтажу [7]. Однак стиснення платформою, повторне кодування, кадрування, знімання екрана, додавання шуму чи поява нової генеративної моделі можуть зменшити надійність детектора. Це дає підстави запропонувати авторське положення про асиметричність результатів автоматизованої детекції: виявлена технічна аномалія може посилювати версію про маніпуляцію, однак невиявлення аномалій саме по собі не підтверджує автентичності матеріалу.

Порівняння наявних підходів визначає місце запропонованої моделі. NIST систематизує технології детекції, автентифікації, маркування й установлення походження синтетичного контенту [8]. C2PA забезпечує криптографічно перевірювані твердження про походження та історію змін активу, проте не встановлює правдивості змісту; відсутність Content Credentials не доводить фальсифікації [9]. Протокол Берклі регламентує збирання, збереження, аналіз і представлення цифрової інформації з відкритих джерел [10]. Авторська модель не створює нового детектора або стандарту походження: вона інтегрує технічний аналіз, контекстну перевірку та документування поводження з файлом у структуру, орієнтовану на правоохоронне реагування і процесуальне оцінювання. Її новизна полягає у визначенні функцій, послідовності та меж застосування цих засобів для формування доказового висновку.

Для правоохоронної практики авторкою пропонується п'ятикомпонентна модель верифікації синтетичного цифрового матеріалу. Вона не є затвердженим стандартом або апробованим протоколом, а становить концептуальну основу, що поєднує п'ять взаємопов'язаних компонентів перевірки: походження, технічні артефакти, контекстну відповідність, цілісність і документовану відтворюваність. Перший компонент – походження – охоплює встановлення першоджерела, автора або пристрою, найранішої доступної публікації, історії передавання, C2PA-маніфесту, метаданих і зв'язку файла з іншими цифровими об'єктами. Ключовим є питання не «чи виглядає запис справжнім», а «яким перевірюваним шляхом він потрапив до дослідника». У воєнних умовах цей шлях може бути фрагментарним, тому кожна прогалина має фіксуватися, а не заповнюватися припущенням.

Другий компонент – технічні артефакти – передбачає мультимодальний аналіз. Для зображень перевіряються освітлення, тіні, перспектива, відбиття, анатомія, текстури та шум; для відео – міжкадрова узгодженість, рух губ, міміка й повторне кодування; для аудіо – спектр, дихання, просодія та ознаки клонування голосу. Результати кількох інструментів зіставляються з урахуванням версії засобу, порогу рішення, навчального набору й можливості роботи з невідомим генератором.

Третій компонент – контекстна відповідність – охоплює зіставлення матеріалу з геолокацією, погодою, освітленням, супут-

никовими знімками, хронологією, офіційними повідомленнями, незалежними відео, свідченнями та даними сенсорів. OSINT перетворює окремий файл на систему перевірюваних зв'язків. Семантична правдоподібність не замінює перевірки: синтетичний матеріал може бути переконливим, а автентичний – містити незвичні деталі бойової обстановки.

Четвертий компонент – цілісність – стосується незмінності та контрольованого обігу даних. Під час первинного одержання зберігається доступний оригінал, обчислюється хеш-значення, створюється захищена контрольна копія, а подальше дослідження виконується на робочій копії. Фіксуються URL, ідентифікатор публікації та профілю, дата, час і часовий пояс, спосіб завантаження, структура каталогу, супровідні повідомлення й контекст сторінки. Кожна операція конвертації, вилучення кадру або покращення аудіо має залишати запис у журналі аудиту. Такий підхід узгоджується з вимогами Протоколу Берклі щодо системного збирання, збереження, аналізу й представлення цифрової інформації з відкритих джерел [10], а також із запропонованим М. Погорецьким тестом «походження – відкриття – відтворюваність» [1].

П'ятий компонент – документована відтворюваність – означає можливість незалежного фахівця повторити процедуру й зрозуміти межі висновку. У звіті зазначаються назва та версія інструменту, параметри аналізу, контрольні файли, отримані показники, інтервали невизначеності, альтернативні пояснення та причини відхилення кожної версії. Категоричний висновок «deepfake/не deepfake» доцільно замінити градуйованим: підтверджено синтетичне походження; виявлено сукупність ознак синтезу; походження не встановлено; ознак синтезу не виявлено, але автентичність не підтверджено; автентичність підтверджено незалежною сукупністю даних. Така шкала не маскує невизначеність і є більш придатною для процесуального оцінювання.

Запропонована модель охоплює три доказові сценарії. У першому матеріал підтверджує зафіксовану подію, тому потребує найвищого рівня встановлення походження та контексту. У другому синтетичний об'єкт є предметом або засобом правопорушення: документуються його створення, властивості, поширення, аудиторія та цифрові зв'язки з причетною особою. У третьому сторона заперечує автентичний матеріал як AI-згенерований; така версія має

містити конкретні перевірювані ознаки, а не посилання на абстрактну можливість deepfake.

Практичну логіку моделі ілюструють два умовні сценарії, які не є результатами емпіричної апробації. У першому в соціальній мережі поширено відеозвернення нібито керівника органу влади із закликом до евакуації. Встановлюються найраніша публікація, обліковий запис, метадані й маршрут поширення; аналізуються синхронізація мовлення, міжкадрові аномалії, спектр голосу та повторне кодування; зміст зіставляється з офіційними каналами, геолокацією, часом і незалежними записами. У другому автентичний відеозапис заперечується стороною, інтересам якої він суперечить. Саме твердження про застосування ШІ не спростовує матеріалу: перевіряються походження файла, безперервність зберігання, метадані, незалежні підтвердження та ознаки втручання. В обох випадках фіксуються хеш-значення, спосіб одержання, версії інструментів і перетворення копії. Отже, модель має не лише виявляти синтетичне, а й захищати автентичне від необґрунтованої дискредитації; негативний результат детектора без підтвердженого походження не є достатнім для остаточного висновку.

В умовах війни застосування запропонованої моделі має доповнюватися оцінюванням часової критичності. Матеріал, здатний спричинити паніку, імітувати наказ, спровокувати вогневе ураження чи приховати воєнний злочин, потребує швидкого попереднього висновку. Проте швидкість не повинна перетворювати попередній результат детектора на остаточний доказ. Доцільно використовувати дворежимну модель: оперативна верифікація формує короткий висновок для негайного реагування із зазначенням рівня впевненості; повна криміналістична верифікація забезпечує збереження оригіналу, повторюваність методики та процесуальне оформлення. Це знижує ризик як запізнілого реагування, так і помилкового публічного спростування автентичного матеріалу.

Наукова новизна запропонованого підходу полягає, по-перше, у розмежуванні синтетичності цифрового об’єкта та його доказової функції; по-друге, в обґрунтуванні авторського положення про асиметричність результатів автоматизованої детекції; по-третє, у концептуалізації переходу від одноканального розпізнавання до комплексної верифікації походження цифрового об’єкта; по-чет-

верте, у розробленні п'ятикомпонентної моделі верифікації, запропонованої як предмет подальшої емпіричної апробації; по-п'яте, у включенні «дивіденду брехуна» до моделі криміналістичних ризиків. Модель має концептуальний характер і є науково обґрунтованою гіпотезою, а не валідованим криміналістичним протоколом; її результативність має бути перевірена на контрольованих наборах і реальних кейсах із відомим походженням матеріалів. Практичним результатом апробації може стати стандартизована картка верифікації, яка міститиме окремі поля для походження, технічних артефактів, контекстної відповідності, цілісності, відтворюваності та рівня висновку.

Таким чином, генеративний ШІ не лише дає змогу створити переконливу підробку, а й дозволяє позбавити справжній запис доказової сили шляхом неконкретизованого сумніву. Протидія має ґрунтуватися на встановленні походження цифрового об'єкта, збереженні його цілісності, мультимодальному аналізі, зовнішній тріангуляції та відтворюваності операцій. Запропонована модель може стати концептуальною основою переходу від імовірнісного машинного сигналу до аргументованого криміналістичного висновку лише після емпіричної валідації. Подальші дослідження мають визначити мінімальну сукупність незалежних підтверджень, перевірити модель на матеріалах різних генераторів і сформувати міжвідомчі стандарти документування синтетичного контенту у провадженнях, пов'язаних зі збройною агресією.

Список використаних джерел:

1. Погорецький М.А. Цифрові технології та докази у розслідуванні злочинів проти основ національної безпеки України: процесуальні проблеми та європейські стандарти. Аналітично-порівняльне правознавство. 2025. № 5, ч. 3. С. 239–255. DOI: https://doi.org/10.24144/2788–6018.2025.05.3.37.

2. Калюжна Ю.І. Інформаційна війна в епоху синтетичних медіа: діпфейки як загроза національній безпеці України та ЄС. Національні інтереси України. 2026. № 5 (22). С. 3316–3372. DOI: https://doi.org/10.52058/3041–1793–2026–5(22)-3316–3372.

3. Chesney R., Citron D.K. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. California Law Review. 2019. Vol. 107, No. 6. P. 1753–1819.

4. Vaccari C., Chadwick A. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social Media + Society. 2020. Vol. 6, No. 1. DOI: https://doi.org/10.1177/2056305120903408.

5. Twomey J., Ching D., Aylett M. P., Quayle M., Linehan C., Murphy G. Do Deepfake Videos Undermine Our Epistemic Trust? A Thematic Analysis of Tweets That Discuss Deepfakes in the Russian Invasion of Ukraine. PLOS ONE. 2023. Vol. 18, No. 10. Article e0291668. DOI: https://doi.org/10.1371/journal.pone.0291668.

6. Mirsky Y., Lee W. The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys. 2021. Vol. 54, No. 1. Article 7. P. 1–41. DOI: https://doi.org/10.1145/3425780.

7. Verdoliva L. Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing. 2020. Vol. 14, No. 5. P. 910–932. DOI: https://doi.org/10.1109/JSTSP.2020.3002101.

8. National Institute of Standards and Technology. Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. NIST AI 100–4. 2024. URL: https://doi.org/10.6028/NIST.AI.100–4 (дата звернення: 13.09.2026).

9. Coalition for Content Provenance and Authenticity. C2PA Specifications. Version 2.4. 2026. URL: https://spec.c2pa.org/ (дата звернення: 13.09.2026).

10. Office of the United Nations High Commissioner for Human Rights; Human Rights Center, University of California, Berkeley. Berkeley Protocol on Digital Open Source Investigations. New York; Geneva, 2022. URL: https://www.ohchr.org/sites/default/files/2022–04/OHCHR_BerkeleyProtocol.pdf (дата звернення: 13.09.2026).