УДК 355.45:316.77
DOI https://doi.org/10.32782/apfs.v054.2025.33

***В. А. Савчук***
*ORCID ID: https://orcid.org/0009-0008-6051-0252*
*аспірант кафедри соціально-гуманітарних дисциплін*
*ПВНЗ «Європейський університет»*

# ІНСТРУМЕНТИ ВИЯВЛЕННЯ ТА ПОДОЛАННЯ ДЕСТРУКТИВНИХ ВПЛИВІВ ТА ПРОПАГАНДИСТСЬКИХ НАРАТИВІВ ІНФОРМАЦІЙНИХ ОПЕРАЦІЙ: ТИПОЛОГІЯ, ОСНОВНІ ХАРАКТЕРИСТИКИ

**Постановка проблеми.** Стрімка еволюція технологій когнітивного впливу в інформаційно-психологічних операціях, які проводить Росія в рамках гібридної війни проти України, створює запит на розробку й застосування ефективних інструментів для виявлення і нейтралізації деструктивного контенту. Проблема має не лише безпековий, а й політико-управлінський вимір: від своєчасності і точності детекції залежить стійкість демократичних інститутів, довіра до владних рішень і здатність держави координувати стратегічні комунікації. У статті обґрунтовується, що арсенал протидії безперервно розширюється та ускладнюється, а отже потребує цілісної типологізації за функціями, ступенем автоматизації та сферою застосування.

**Аналіз останніх досліджень і публікацій та виокремлення невирішених частин проблеми.** Питання розробки і впровадження ефективних інструментів виявлення дезінформації в медіапросторі різнопланово досліджується в роботах як зарубіжних, так і українських науковців. Ключовою для технічної лінії джерел є робота групи дослідників Університетів Падуї та Болоньї, які у 2019 році представили першу загальнодоступну систему для виявлення пропаганди в онлайн-новинах в режимі реального часу Proppy. На думку автора, впровадження в практику цієї системи заклало основу для подальших автоматизованих підходів до детекції риторичних прийомів і маніпуляцій у ЗМІ. Поряд із західними підходами згадуються й новітні китайські розробки, зокрема розглянутий у дослідженні науковців Університет Яньшань алгоритм BotSAI, що інтегрує мультимодальні ознаки (текст, мережеві зв'язки, метадані) для точнішого розрізнення ботів, у т.ч. тих, що маскуються під «звичайних» користувачів. Другий кластер публікацій стосується донастроєних трансформерів (зокрема RoBERTa) для автоматичної класифікації «правда/фейк» у новинах. Стаття посилається на емпіричні результати команди Університету інженерії та технологій Лахора. До «адаптивного шару» автор відносить великі мовні моделі (LLM) у zero-shot-режимі та, спираючись на результати дослідження групи під керівництвом професорки Людмили Заволокіної з Університет Лозанни, доходить висновку, що додавання покрокового міркування (chain-of-thought) підвищує якість і прозорість верифікації для регуляторів та експертів. Розглядаючи нормативно-дискурсивний вимір автор використовує аналітичні огляди Reporters Without Borders (RSF) та PolitiFact щодо феномену *pseudo-fact-checking* (зокрема кейсу «War on Fakes»), які демонструють, як «фактчекерська» риторика використовується для ретрансляції прокремлівських тез і створення ілюзії нейтральності перевірки. Відзначаючи значний прогрес у розвитку інструментарію виявлення та подолання деструктивного впливу в інфопросторі, автор підкреслює наявність розриву між поширенням терміна в публічному дискурсі та станом внутрішньої нормативної бази України, яка перебуває на етапі концептуального узгодження з міжнародними підходами. Невирішеною лишається проблема формалізації національних індикаторів і метрик когнітивної безпеки

**Мета статті.** Сформувати функціонально орієнтовану типологію інструментів виявлення та подолання деструктивних впливів і пропагандистських наративів, які поширює Росія в рамках гібридної агресії проти України. Описати їх технологічні засади, сфери застосування та емпіричні результати апробації; запропонувати багатошарову архітектуру протидії для політичної практики (алгоритмічна детекція → мережевий моніторинг інокуляція → фактчекінг).

**Виклад основного матеріалу.** Сучасна типологія інструментів протидії дезінформації формується на стику технологічних інновацій, медіаосвіти, критичного аналізу й міждисциплінарних методик. Арсенал інструментів безперервно розширюється та ускладнюється, що потребує системного осмислення і наукової класифікації. В рамках даного дослідження спробуємо виокремити чотири основні типи інструментів протидії деструктивному інформаційному впливу, що відрізняються за цілями, алгоритмами функціонування та рівнем втручання в комунікаційне середовище.

1. Інструменти автоматичного виявлення пропагандистського та дезінформаційного контенту.

Ця категорія охоплює системи, що використовують алгоритми машинного навчання (НЛП, трансформери, BERT-моделі, графові нейромережі) для аналізу тексту, виявлення пропагандистських технік, маніпуляцій, наративів. Застосування фрагментарного анотованого корпусу та багаторівневих моделей дозволяє ідентифікувати специфічні стилістичні прийоми та маркери дезінформації не лише в новинах, поширюваних засобами масової інформації, а й у соціальних мережах. Перша загальнодоступна система виявлення пропаганди в реальному часі для онлайн-новин Proppy була представлена ще в 2019 році групою вчених з університетів Падуї та Болон'ї [1]. Вона спрямована на підвищення обізнаності, потенційно обмежуючи вплив пропаганди та допомагаючи боротися з дезінформацією. Система постійно відстежує низку джерел новин, видаляє дублікати та кластеризує новини за подіями, а також організовує статті про подію на основі ймовірності того, що вони містять пропагандистський контент. На сьогодні, найпростіший рівень цієї екосистеми становлять класичні трансформер-класифікатори на кшталт донавченої моделі RoBERTa (робастно оптимізованої версії BERT, додатково налаштованої на спеціалізованих иасивах «правда/фейк» для підвищення доменної чутливості), які автоматично маркують повідомлення за стилістичними та семантичними ознаками [2]. Дослідження науковців з Університету інженерії та технологій Лахора підтвердили, що на «чистих» корпусах новин RoBERTa демонструє майже ідеальні показники: 99,99 % точності на наборі ISOT (набір даних про фейкові новини ISOT — це збірка кількох тисяч фейкових новин та правдивих статей, отриманих з різних легітимних новинних сайтів та сайтів, позначених сайтом Politifact.com як ненадійні) і 86,16 % на FakeNewsNet [3]. Другий – «адаптивний» – шар утворюють великі мовні моделі (LLM), які працюють у zero-shot-режимі («навчання без прикладів» має на увазі здатність штучного інтелекту освоювати нові завдання без навчання на конкретних прикладах) й дозволяють оперативно оцінювати сумнівний контент навіть без попереднього маркування. Однак, емпірична валідація у дослідженні групи швейцарських вчених під керівництвом професора Університету Лозанни Людмили Заволокіної засвідчує: базовий запит «Чи є це дезінформацією?» дає істотно гірші результати, ніж класичний детектор, тоді як використання ланцюга міркувань (Chain-of-Thought) суттєво підвищує точність і робить процес прозорим для регуляторів та експертів [4]. Універсалізація великих мовних моделей (LLM) та донавчених трансформерів, як-от модуль ClarifAI, формує новий шар «цифрового суверенітету» громадянина: аналіз здійснюється локально у браузері, а візуальні маркер-попередження *(stumbling stones)* та короткі пояснення (*digital nudges*) переводять споживання новини з автоматизованої Системи 1 (за Д. Канеманом) у рефлексивну Систему 2, підвищуючи критичність аудиторії з 82 % до 96 % і збільшуючи середній час обмірковування повідомлення майже на хвилину.

2. Системи моніторингу мережевого поширення та бот-активності. Масовість аудиторій, швидкість розповсюдження контенту й можливість автоматизованої взаємодії зробили сучасні соціальні мережі Twitter/X, Telegram, Facebook, TikTok ключовими аренами інформаційно-психологічного впливу. У цьому контексті системи виявлення бот-активності стали критично необхідними для своєчасної діагностики, локалізації та нейтралізації інформаційних загроз в соцмережах. Такі інструменти дозволяють ідентифікувати скоординовану поведінку облікових записів, виявляти штучне розповсюдження пропагандистських меседжів і моделювати мережеву структуру маніпулятивного впливу. Першим безкоштовним та відкритим для громадськості програмним забезпеченням, що сканує соціальні мережі в режимі реального часу, щоб виявити докази автоматизованих акаунтів у Twitter (ботів, які скоординовано розсилають повідомлення), був BotSlayer [5]. Цей інструмент для боротьби з дезінформацією в Інтернеті, створений в Обсерваторії соціальних мереж Університету Індіани (США), надав спільнотам та окремим користувачам можливість виявляти скоординовані дезінформаційні кампанії без будь-якого попереднього знання про ці кампанії. Більш сучасним інструментом є BotSAI – створена в серпні 2024 року групою науковців з Університету Яньшань (КНР) AI-система для виявлення бот-акаунтів у Twitter (зараз X). У платформі інтегруються мультиканальні ознаки користувачів (текст, мережеві властивості, користувацькі метадані) для точного розпізнавання ботів, включаючи тих, хто маскується під реальних юзерів [6]. Також в квітні 2024 року науковці з Бейханського університету (КНР) представили UnDBot, нову платформу для виявлення соціальних ботів, побудовану на теорії структурної інформації. Модель, створена на основі глибокого машинного навчання, аналізує три метрики соціальних відносин, які враховують різні аспекти поведінки соціальних ботів: розподіл типів публікацій, вплив публікацій та співвідношення користувачів до підписників [7]. PhD Антуан Вастель, провідний розробник інфраструктури для виявлення ботів і зловживань в соцмережах з Лілльського Університету (Франція), у своєму дослідженні стверджує [8], що для ефективної детекції потрібна комплексна оцінка кількох типів метрик, які можна систематизувати за такими категоріями:

a) метрики відбитків (Fingerprinting signals);

b) метрики поведінкової активності (Behavioral signals);

c) метрики репутації (Reputational signals);

d) метрики контексту та мультифакторна оцінка (кореляція контекстних сигналів і слабких сигналів).

Такий багатошаровий підхід, застосований на базі машинного навчання, є однією з найефективніших стратегій для точного і своєчасного виявлення ботів і аномальної активності в соцмережах сьогодні.

3. Інструменти превентивної інокуляції (prebunking) та освітньої протидії. Хоча вже розглянуті інструменти забезпечують оперативне виявлення дезінформаційного контенту та каналів його поширення в режимі реального часу, ключовою залишається проблема моменту взаємодії – більшість таких інструментів активуються лише після створення і дистрибуції деструктивного повідомлення, коли його вплив на цільову аудиторію вже відбувся або розпочався. Саме в цьому контексті набувають особливої значущості інструменти превентивної інокуляції (*prebunking*), які спрямовані не на постфактум реагування, а на запобігання самому акту когнітивного впливу. Інструменти prebunking-у і освітньої протидії базуються на класичній теорії психологічної інокуляції, запропонованій соціальним психологом Вільямом Макгвайром у 1960-х роках. За аналогією з біологічною імунізацією, суть підходу полягає в тому, що сприйняття людини з ослабленими формами маніпуляцій – зразками фейків, типів прийомів чи дезінформаційних стратегій – формує у ній стійкість до майбутніх інформаційних атак. С. ван дер Лінден, дослідник з Кембріджського Університету (Велика Британія), стверджує, що попередні повідомлення, засновані на цій базовій структурі, «можуть посилити психічну стійкість глядача до спроб переконання, подібно до того, як медичні вакцини створюють фізіологічний опір майбутнім інфекціям» [9]. Практична реалізація теорії інокуляції складається з кількох етапів. По-перше, людина отримує попередження про можливість маніпуляцій, що «запускає психологічну імунну систему і підвищує скептицизм та настороженість». По-друге, вона знайомиться з характерними тактиками фейків у м'якій формі (наприклад, у навчальній грі чи відео), що дає змогу заздалегідь підготувати контраргументи та розвинути критичне мислення [10]. Експерти з незалежної дослідницької і доказової агенції Verian Group (Лондон) у 2023 провели масштабний онлайн- експеримент у Великій Британії, в ході якого оцінили значний статистичний вплив превентивної інокуляції на зниження вразливості до дезінформаці. Результати дослідження показали, що учасники, які пройшли інокуляційне навчання, «лайкали» або «любили» фейкові пости вдвічі менше за учасників контрольної групи [11].

4. Фактчекінгові платформи і цифрові системи верифікації. Індивідуальна обізнаність, здатність до розпізнавання маніпулятивних технік та усвідомлене споживання контенту залишаються незамінними компонентами системної протидії дезінформації. Сучасні фактчекінгові платформи і цифрові засоби перевірки надають кінцевому користувачу реальні можливості здійснювати перевірку контенту самостійно. Один із найпопулярніших у світі порталів перевірки фактів в інтернеті Snopes розпочав свою діяльність ще у 1994 році, його методи аналізу і рейтинги оцінки інформації стали своєрідним стандартом у сфері фактчекингу. Кожна перевірка тверджень, які аналізує Snopes оцінюється за шкалою достовірності, що налічує 20 рівнів [12]. Наприклад, якщо основні елементи твердження є очевидно правдивими, але деякі допоміжні деталі, що стосуються твердження, можуть бути неточними, йому надається рейтинг «Здебільшого правда». Твердження, стосовно якого не знайдено переконливих доказів про його істинність (такі твердження зазвичай виникають з чуток, спекуляцій або безпідставних чуток), маркується як «Безпідставний». Міжнародна мережа перевірки фактів IFCN, запущена у 2015 році Інститут вивчення медіа Пойнтера (Poynter Institute, США), встановила кодекс етики [13] для організацій, що здійснюють перевірку фактів. IFCN на сьогодні виконує роль стандартизатора та координатора у глобальній боротьбі з дезінформацією. Співробітництво між IFCN і великими соціальними платформами, такими як Facebook і TikTok, сприяє впровадженню механізмів маркування та зменшенню зростання неправдивого контенту. Станом на липень 2024 року IFCN налічувала 170 організацій -членів [14]. В Україні з 2014 року успішно функціонує платформа StopFake, яка декларує стратегію MAD (моніторинг, архівування, розвінчання) і працює за принципами Кодексу Міжнародної мережі перевірки фактів (IFCN). З 2023 року організація зосереджена на спростуваннях, пов'язаних з російсько-українською війною [15]. Ще один підписант Кодексу етики Міжнародної мережі фактчекерів Інституту Poynter український проєкт VoxCheck щомісяця додає близько 100 спростованих фейків до власної бази [16], яка налічує понад 6000 спростувань за весь час роботи. Високий рівень прозорості методології (public methodology) і відкритий архів перевірок формують «соціальний капітал довіри» фактчекінгових платформ, який дедалі активніше використовується урядами, медіа та соціальними мережами для оперативного стримування інфор-

маційних атак. Водночас саме цей капітал стає ціллю російських зловмисників, які прагнуть не лише оминати фактчек, а й позбавити інститут перевірки морального авторитету, вдаючись до тактики «імітаційного фактчекінгу». Найяскравішим прикладом є псевдо-глобальна мережа Global Fact-Checking Network (GFCN), анонсована 20 листопада 2024 р. на форумі «Dialogue on Fakes 2.0» і просунута під егідою кремлівської АНО «Диалог» та агентства ТАСС [17]. За оцінкою Reporters Without Borders, GFCN з квітня 2025 р. підтримує сайт англійською й російською, декларує 55 «експертів» із 37 країн, але фактично поширює прокремлівські наративи й імітує бренд IFCN, щоб «послабити авторитет незалежних перевірчих мереж» [18]. До цієї ж групи належить і сайт War on Fakes (створений 2022 р., активний у 2024–2025 рр.), який копіює стиль фактчек-статей для виправдання російської агресії проти України [19]. Російські псевдо-фактчекінгові ініціативи не просто продукують хибний контент, вони розмивають епістемічні межі між достовірним та авторитетним, підриваючи довіру до інструментів перевірки загалом. На фоні вибухового розвитку систем штучного інтелекту (ШІ) все більшого поширення набувають цифрові системи верифікації. Метадані, комп'ютерний зір і глибокі нейронні мережі дають змогу автоматизувати раннє виявлення маніпулятивного контенту, скорочуючи час реакції користувачів. На рівні агрегованого фактчекінгу Google Fact Check Explorer [20] і Fact Check Markup Tool [21] демонструють, як ШІ допомагає структурувати мільйони записів ClaimReview і робити їх доступними через єдиний пошуковий інтерфейс. Завдяки розмітці ClaimReview у схемах структурованих даних ці інструменти інтегрують незалежні перевірки в основний індекс Google Search і автоматично підтягують релевантні фактчек-картки до запитів користувачів [22]. Платформа Deepware Scanner пропонує веб-інтерфейс, що оцінює ймовірність синтетичних змін у відео за допомогою згорткових мереж, орієнтуючись на журналістів і правоохоронців [23]. Для перевірки статичних зображень та відео-метаданих використовуються TinEye і YouTube Data Viewer. Перший індексує понад 70 млрд зображень і застосовує власні алгоритми візуального «відбитка» для пошуку навіть сильно відредагованих копій, що робить його стандартним інструментом у редакційних «шейт-листах» (списках джерел, яких треба уникати або використовувати з особливою обережністю) журналістської верифікації [24]. Другий, розроблений Amnesty International, автоматично видобуває приховані таймстемпи та генерує прев'ю-кадри з роликів YouTube [25]. Сучасний інструментарій цифрової верифікації переходить від ізольованих модулів до багаторівневих ШІ-екосистем, де різнотипні алгоритми взаємодоповнюються. Втім, як підтверджують дослідження науковців з Університету Бергена (Норвегія) людський досвід все ще є критично важливим для перевірки фактів за допомогою штучного інтелекту, оскільки системи штучного інтелекту не можуть повністю зрозуміти контекст, намір чи достовірність [26]. Однією з невирішених на сьогодні проблем є кодування складних понять, таких як етика та критичне мислення, в алгоритми ШІ.

**Висновки і перспективи подальших досліджень.** У підсумку стаття доводить, що сучасний інструментарій протидії дезінформації є багаторівневим і охоплює як реактивні, так і превентивні механізми. З огляду на еволюцію процесу, технологічна конвергенція (НЛП / LLM/ комп'ютерний зір/ мережевий аналіз) зумовлює перехід до екосистемного підходу, який поєднує всі фактори. Водночас, автор робить висновок, що людська експертиза лишається незамінною ланкою, зокрема для інтерпретації наміру, контексту та хронології подій. Подальші дослідження мають бути спрямовані на підвищення прозорості стандартів, обов'язковий аудит учасників перевірочних мереж, розвиток медіа-грамотності користувачів та розв'язання етичних питань при розробці і впровадженні алгоритмів ШІ.

***Література***

1. Da San Martino G., Barrón-Cedeño A., Petroni F., et al. Proppy: A System to Unmask Propaganda in News. arXiv preprint. 2019. URL: https://arxiv.org/abs/1912.06810 (дата звернення: 23.05.2025).

2. Liu Y., Ott M., Goyal N., та ін. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint. 2019. arXiv:1907.11692 [cs.CL]. 15 с. URL: https://arxiv.org/abs/1907.11692 (дата звернення: 23.05.2025).

3. Maham S., Uzir N., Rahman Z., et al. ANN: Adversarial News Net for Robust Fake News Classification. IEEE Access. 2024. Vol. 12. С. 55890–55907.

4. Zavolokina L., et al. ClarifAI: A Multilingual AI System for Propaganda Prebunking. arXiv preprint. 2024. URL: https://arxiv.org/abs/2402.19135 (дата звернення: 13.06.2025).

5. Tracking coordinated disinformation campaigns online made easier with new BotSlayer tool [Електронний ресурс]. Indiana University. URL: https://news.iu.edu/live/news/26311-tracking-coordinated-disinformation-campaigns (дата звернення: 13.06.2025).

6. Gong J., Li Z., Wang H. Enhancing Twitter Bot Detection via Multimodal Invariant Representation (BotSAI). arXiv preprint. 2024. С. 1–2. URL: https://arxiv.org/abs/2408.03096 (дата звернення: 15.06.2025).

7. Peng H., Zhang J., Huang X., et al. Unsupervised Social Bot Detection via Structural Information Theory. arXiv preprint. 2024. URL: https://arxiv.org/abs/2404.13595 (дата звернення: 15.06.2025).

8. Bot detection 101: How to detect bots In 2025? URL: https://blog.castle.io/bot-detection-101-how-to-detect-bots-in-2025-2 (дата звернення: 16.06.2025).

9. Van der Linden S. Countering misinformation through psychological inoculation. *Inoculation Science*. 2024. С. 2–3. URL: https://inoculation.science (дата звернення: 16.06.2025).

10. Roozenbeek J., van der Linden S. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*. 2022. С. 5. URL: https://www.science.org/doi/10.1126/sciadv.abo6254 (дата звернення: 16.06.2025).

11. McPhedran R., Ratajczak M., Mowby M., et al. Psychological Vaccination Protects Against Social Media Infodemic. *Scientific Reports*. 2023. Вип. 13:5780. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10082776/ (дата звернення: 04.06.2025).

12. Snopes. Fact-check ratings. URL: https://www.snopes.com/fact-check-ratings/ (дата звернення: 19.06.2025).

13. International Fact-Checking Network. The commitments of the Code of Principles. URL: https://ifcncodeofprinciples.poynter.org/the-commitments (дата звернення: 19.06.2025).

14. Duke Reporters' Lab. Fact-checking Census 2024. URL: https://www.poynter.org/fact-checking/2024/duke-reporters-lab-fact-checking-cencus-2024 (дата звернення: 05.06.2025).

15. StopFake. 5 проєктів по боротьбі з дезінформацією. URL: https://www.stopfake.org/uk/5-proyektiv-po-borotbi-z-dezinformatsiyeyu/ (дата звернення: 05.06.2025).

16. VoxCheck. Платформа фактчекінгу в Україні. URL: https://voxcheck.voxukraine.org/ (дата звернення: 05.06.2025).

17. EDMO. Putin's new plan to undermine fact-checking. 29 листоп. 2024. URL: https://edmo.eu/publications/putins-new-plan-to-undermine-fact-checking (дата звернення: 05.06.2025).

18. Reporters Without Borders. Russia: fact-checking is the Kremlin's latest propaganda tool. URL: https://rsf.org/en/russia-fact-checking-kremlins-latest-propaganda-tool (дата звернення: 07.06.2025).

19. PolitiFact. How 'War on Fakes' uses fact-checking to spread pro-Russia propaganda. URL: https://www.politifact.com/article/2022/aug/08/how-war-fakes-uses-fact-checking-spread-pro-russia/ (дата звернення: 05.06.2025).

20. Google. Fact Check Explorer. 2025. URL: https://toolbox.google.com/factcheck/explorer (дата звернення: 05.06.2025).

21. Google. Fact Check Markup Tool. 2025. URL: https://toolbox.google.com/factcheck/markuptool (дата звернення: 07.06.2025).

22. Google Developers. Fact Check (ClaimReview) structured data. 2025. URL: https://developers.google.com/search/docs/appearance/structured-data/factcheck (дата звернення: 07.06.2025).

23. Deepware. Scan & Detect Deepfake Videos. 2025. URL: https://deepware.ai (дата звернення: 07.06.2025).

24. TinEye. Reverse Image Search Engine. Берез. 2025. URL: https://en.wikipedia.org/wiki/TinEye (дата звернення: 07.06.2025).

25. YouTube Data Viewer – Citizen Evidence Lab. Amnesty International. 2014 (оновл. 2025). URL: https://citizenevidence.amnestyusa.org (дата звернення: 07.06.2025).

26. EDMO. Part of the problem and part of the solution: the paradox of AI in fact-checking. 2025. URL: https://edmo.eu/blog/part-of-the-problem-and-part-of-the-solution-the-paradox-of-ai-in-fact-checking (дата звернення: 08.06.2025).

**Анотація**

***Савчук В. А.*** **Інструменти виявлення та подолання деструктивних впливів та пропагандистських наративів інформаційних операцій: типологія, основні характеристики.** – Стаття.

Стаття системно аналізує інструменти виявлення й подолання деструктивних впливів та пропагандистських наративів у сучасних інформаційних операціях. Мета статті – класифікувати засоби протидії за ступенем автоматизації (ручні, автоматизовані, гібридні), технологічною основою (NLP, AI, OSINT, фактчек), сферою застосування (текст, візуальні/аудіо-матеріали, соцмережі) та функціями (детекція, профілактика, контрнаративи). Методологія дослідження поєднує порівняльний огляд рішень і узагальнення емпіричних даних цифрових кампаній. Перший кластер це алгоритмічна детекція. Трансформери й графові нейромережі ідентифікують риторичні техніки, семантичні патерни та зв'язки, приклад Propy і донавчені моделі на кшталт RoBERTa. LLM у zero-shot-режимі забезпечують швидку оцінку, а прозорі інтерфейси підвищують підзвітність. Другий мережевий моніторинг і виявлення скоординованої поведінки (BotSlayer, BotSAI, UnDBot), що інтегрує текстові, топологічні та метадані для картографування каналів поширення. Третій – превентивна інокуляція (prebunking) й освітні інтервенції, які зменшують вразливість до фейків незалежно від рівня когнітивної рефлексії. Четвертий – фактчекінг і цифрова верифікація (стандарти IFCN, інструменти виявлення deepfake та зворотного пошуку зображень). Наукова новизна полягає у функціонально орієнтованій типології інструментів з інтеграцією технічних, соціально-психологічних і комунікативних вимірів. Практична значущість полягає у пропозиції багатошарової архітектури протидії: алгоритмічна детекція → мережевий моніторинг → інокуляція → фактчекінг. Ключовий висновок, який робить автор у своєму дослідженні: ефективність виявлення деструктивних впливів і дезінформації забезпечує гібридна модель human-in-the-loop, де автоматизовані системи невіддільні від експертизи, етичних стандартів і прозорих процедур аудиту.

*Ключові слова:* дезінформація, великі мовні моделі, бот-мережі, мережевий моніторинг, інокуляція, факт-чекінг, цифрова верифікація.

**Summary**

***Savchuk V. A.*** **Tools for identifying and overcoming destructive influences and propaganda narratives in information operations: typology, key characteristics.** – Article.

The article offers a systematic analysis of instruments for detecting and countering destructive influences and propagandist narratives in contemporary information operations. The aim is to classify countermeasures by degree of automation (manual, automated, hybrid), technological basis (NLP, AI, OSINT, fact-checking), sphere of application (text, visual/audio materials, social networks), and function (detection, prevention, counter-narratives). The research methodology combines a comparative review of solutions and a synthesis of empirical evidence from digital campaigns. The first cluster is algorithmic detection. Transformers and graph neural networks identify rhetorical techniques, semantic patterns, and relationships; examples include Proppy and fine-tuned models such as RoBERTa. Large language models in zero-shot mode provide rapid assessment, while transparent interfaces increase accountability. The second is network monitoring and the detection of coordinated behavior (BotSlayer, BotSAI, UnDBot), which integrates textual, topological, and metadata signals to map dissemination channels. The third comprises preventive inoculation (prebunking) and educational interventions that reduce susceptibility to falsehoods regardless of the level of cognitive reflection. The fourth concerns fact-checking and digital verification (IFCN standards, deepfake-detection tools, and reverse image search). The study's scholarly novelty lies in a function-oriented typology of instruments that integrates technical, socio-psychological, and communicative dimensions. Its practical significance is a multilayered architecture of counteraction: algorithmic detection → network monitoring → inoculation → fact-checking. The key conclusion drawn by the author is that effective detection of destructive influence and disinformation is ensured by a human-in-the-loop hybrid model, in which automated systems are inseparable from expert judgment, ethical standards, and transparent audit procedures.

*Key words:* disinformation, large language models, bot-networks, network monitoring, inoculation, fact-checking, digital verification.

Дата надходження статті: 25.06.2025
Дата прийняття статті: 07.07.2025
Опубліковано: 10.09.2025