ПРИКЛАДНІ СОЦІАЛЬНО-КОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ

УДК 316.774:004.738.5:070
DOI https://doi.org/10.32782/2710-4656/2026.3.2/28

Баловсяк Н. В.
https://orcid.org/0000-0002-1810-7397
Заклад вищої освіти «Український католицький університет»

БОРОТЬБА З ФЕЙКОВИМ КОНТЕНТОМ У СОЦІАЛЬНИХ МЕРЕЖАХ У 2015–2022 РОКАХ: ІСТОРИЧНИЙ АСПЕКТ

У статті здійснено комплексний історичний аналіз підходів соціальних мереж до протидії поширенню фейкового контенту в період 2015–2022 років – до появи та масового застосування технології генеративного штучного інтелекту. Дослідження охоплює трансформацію стратегій боротьби з дезінформацією від хаотичних експериментів окремих платформ до відносно системних, але все ще фрагментарних механізмів модерації. Проаналізовано ключові етапи формування політик соціальних мереж щодо боротьби з фейками: наслідки президентських виборів у США 2016 року, скандал із компанією Cambridge Analytica 2018 року, пандемію COVID-19 та події навколо виборів 2020 року в США. Систематизовано основні інструменти протидії дезінформації: ручну модерацію, партнерства з фактчекерами, алгоритмічне зниження видимості проблемного контенту, маркування маніпулятивних матеріалів та виявлення скоординованих інформаційних операцій. Окремо розглянуто еволюцію кожного з цих інструментів у часі, їхні переваги та структурні обмеження в умовах зростаючих обсягів користувацького контенту. Розглянуто специфіку підходів найпопулярніших платформ щодо боротьби із фейками – Facebook, Twitter і YouTube – та здійснено порівняльний аналіз обраних ними стратегій модерації.

Особливу увагу приділено системним обмеженням обраних моделей протидії дезінформації, зокрема конфлікту між комерційними інтересами платформ і суспільним запитом на достовірність інформації. Досліджено питання прозорості та підзвітності платформ перед суспільством і регуляторами, а також дискусії навколо меж допустимого втручання приватних компаній у публічний інформаційний простір. Доведено, що реактивний характер та недостатня масштабованість тодішніх підходів, а також поява загальнодоступних інструментів генеративного штучного інтелекту (ШІ) після 2022 року створили передумови для переходу до автоматизованих ШІ-орієнтованих рішень у сфері виявлення та нейтралізації дезінформації.

Ключові слова: *фейковий контент, дезінформація, соціальні мережі, модерація контенту, фактчекінг, інформаційні операції, штучний інтелект, гібридні загрози.*

Постановка проблеми. Стрімкий розвиток соціальних мереж та їх перетворення на домінуючий канал споживання новин сформували якісно нову реальність, у якій швидкість поширення контенту радикально перевищує можливості його верифікації. В цих умовах фейковий контент перетворився на інструмент впливу на суспільну свідомість, а також на засіб підриву довіри до демократії та ведення інформаційної й гібридної війни.

Особливої гостроти ця проблема набуває для України, яка з 2014 року перебуває в епіцентрі систематичних інформаційних операцій з боку Росії. Досвід України засвідчив, що інформаційна агресія може як передувати конвенційному збройному нападу, так і супроводжувати його. Окрім того, скоординовані інформаційні операції здатні суттєво впливати на сприйняття конфлікту як всередині країни, так і на міжнародній арені.

Алгоритми ранжування, оптимізовані на максимізацію залучення користувачів, створили сприятливе середовище для вірусного поширення сенсаційного та недостовірного контенту. Таким чином, алгоритмічна логіка платформ і загроза дезінформації виявилися структурно переплетеними: будь-яке втручання в поширення недостовірного контенту неминуче зачіпало й легітимний публічний дискурс. Саме тому платформи були змушені балансувати між захистом інформаційного простору та свободою слова, між швидкістю реагування та точністю рішень, між комерційними інтересами та суспільною відповідальністю. Регулятори, у свою чергу, систематично відставали від розвитку технологій і тактик дезінформації.

Поява після 2022 року генеративних моделей штучного інтелекту (надалі – ШІ) спростила продукування фейкового контенту, що, в свою чергу, потребувало нових підходів до ідентифікації фейків. Саме тому осмислення досвіду боротьби з фейковим контентом у 2015–2022 роках є важливим як для теоретичного розуміння феномену дезінформації, так і для розробки ефективних кроків у сфері медіаполітики та регулювання платформ.

Аналіз останніх досліджень і публікацій. Формування наукового дискурсу навколо проблеми фейкового контенту відбувалося паралельно із усвідомленням суспільством масштабів цього явища. До 2016 року дослідження цифрової дезінформації не виокремлювалися в самостійний напрям. Переломним став 2017 рік, коли Клер Вордл та Хосейн Дерахшан запропонували «концепцію інформаційного розладу» та визначили види фейкового контенту – помилкову інформацію (misinformation), дезінформацію (disinformation) та шкідливу інформацію (mal-information) [32, с. 37]. Того ж року Хант Алкотт та Метью Гентзков здійснили масштабний емпіричний аналіз поширення фейків під час президентських виборів у США. Головним результатом цього дослідження була домінуюча роль соціальних мереж як основного каналу поширення дезінформації [4, с. 217].

Знаковим для розуміння механізмів поширення дезінформації стала робота дослідників із Масачусетського технологічного інституту (MIT) [31, с. 1149], яка на масиві понад 126 тисяч дописів із Twitter доводить, що фейкові новини поширюються у шість разів швидше за достовірний контент. При цьому вони охоплюють на 70% більшу аудиторію. Когнітивний вимір проблеми поширення фейків досліджували Е. Тандос, Ж. В. Лім та Р. Лінг, акцентуючи увагу на ролі упереджень підтвердження (confirmation bias) та евристики доступності (availability heuristic) у сприйнятті недостовірної інформації [26, с. 141]. Вагомий внесок в вивчення боротьби із фейками також зробили дослідницькі організації (Atlantic Council DFRLab, Stanford Internet Observatory, Graphika), які вивчають кросплатформенні інформаційні операції.

В Україні дослідження інформаційного виміру гібридної війни розпочалися ще з 2014 року: Г. Почепцов систематизував механізми інформаційного впливу [2, с. 125], а В. Горбулін концептуалізував інформаційну складову гібридного протистояння [3, с. 189]. Проблематику маніпулятивних технологій і державної інформаційної політики досліджували також Б. Парахонський, Г. Яворська та інші вчені.

Попри значний масив напрацювань, у дослідженнях цього періоду залишається низка невирішених проблем. Більшість емпіричних досліджень зосереджена на окремих платформах або конкретних інформаційних епізодах, наслідком чого є фрагментарна картина без цілісного розуміння галузевих тенденцій. Закритість алгоритмів платформ і обмеженість дослідницького доступу до їхніх даних після скандалу із Cambridge Analytica ускладнили незалежний аудит систем модерації. Недостатньо розробленими залишаються порівняльний аналіз ефективності модерації між платформами, країнами й типами загроз, а також довгострокові суспільні ефекти дезінформації.

Постановка завдання. Метою статті є аналіз еволюції підходів соціальних мереж до протидії фейковому контенту в період до масової доступності інструментів штучного інтелекту та оцінка ефективності цих підходів в історичній перспективі. Завданнями дослідження є аналіз наукових підходів до визначення понять «фейковий контент», «дезінформація» та «місінформація» у контексті соціальних мереж 2015–2022 років; опис ключових етапів формування політик соціальних мереж щодо протидії фейкам у досліджуваний період; систематизація основних інструментів боротьби з фейковим контентом, що застосовувалися різними платформами; аналіз обмежень ефективності цих інструментів в умовах масштабування інформаційних потоків; визначення структурних та інституційних чинників, що впливали на вибір платформами стратегій боротьби з фейковим контентом; окреслення передумов переходу соціальних мереж до ШІ-орієнтованих підходів після 2022 року.

Виклад основного матеріалу. До 2016 року термін «фейкові новини» вживався переважно

для позначення сатиричних матеріалів або очевидних фальсифікацій контенту. Проте президентські вибори в США 2016 року та попередній досвід систематичних інформаційних операцій Росії проти України – з фейками про «розп'ятого хлопчика», сфальсифікованими відео з Донбасу та наративами про «громадянську війну» – наочно показали, як соціальні мережі можуть використовуватися для поширення фейків з метою впливу на фундаментальні демократичні процеси. Академічна спільнота зафіксувала обмеження терміну «fake news» і запропонувала точнішу концепцію «інформаційного розладу» (Вордл і Дерахшан), яка розрізняє неправдиву інформацію, дезінформацію та неправильну інформацію – залежно не тільки від правдивості контенту, а й від намірів його поширення [32, с. 37]. Саме цей контекст змусив платформи переосмислити природу загрози: дезінформація виявилася не сукупністю окремих фейків, а частиною скоординованих операцій [6], через що соціальні мережі перейшли від видалення окремих матеріалів до блокування цілих мереж скоординованої неавтентичної поведінки [10].

Кількісні дослідження цього періоду дозволяють оцінити реальний масштаб інформаційного впливу соціальних мереж. Дослідження Алкотта та Гентцкова показало, що 41,8% переходів на сайти з фейковим контентом відбувалося через соцмережі, тоді як через пошукові системи – лише 22% [4, с. 224]. Показовим є порівняння охоплення: топ-20 найпопулярніших фейкових новин отримав понад 8,7 млн реакцій у Facebook, тоді як публікації провідних новинних агентств – лише 7,3 млн. Це відбувалося на тлі зростаючої залежності аудиторії від соціальних мереж як основного джерела інформації: у 2017 році 67% американців отримували новини саме звідти, хоча довіра до цього каналу була значно нижчою, ніж до традиційних медіа [21]. Цією вразливістю скористалися зовнішні актори: пов'язаний із Росією контент охопив понад 126 млн американців, а витрати на політичну рекламу у Facebook під час виборів перевищили 100 тис. доларів [24].

Реакція Facebook на виявлені проблеми була несистемною й експериментальною. У квітні 2017 року запроваджено механізм маркування спірного контенту [14]: матеріали передавалися незалежним фактчекерам (PolitiFact, Snopes, AP, ABC News, FactCheck.org), а підтверджені фейки позначалися як «disputed» («спірний»). Однак цей інструмент мав суттєвий побічний ефект: користувачі схильні були сприймати непозначений контент як перевірений і достовірний [15], що підвищувало довіру до решти фейків. Найбільш парадоксальним виявився психологічний ефект: дослідження засвідчили, що маркування лише частини матеріалів як спірних підвищувало сприйняту достовірність непозначеного контенту, навіть коли він був хибним – так званий «ефект імпліцитної правди» (implied truth effect) [18]. На початку 2018 року Facebook відмовився від цієї практики, замінивши її на показ пов'язаних матеріалів («related articles») [17, с. 4951]. Паралельно Facebook тестував Explore Feed – окрему стрічку для публікацій від медіа та брендів, яка мала на меті відокремити ці матеріали в окрему вкладку, аби алгоритм основної стрічки орієнтувався на дописи від друзків. Однак експеримент по запровадженню Explore Feed у низці країн провалився: зокрема, у Словаччині медіа втратили до 75% органічного охоплення [25], після чого від глобального впровадження цієї функції відмовилися.

Ці експерименти виявили три ключові закономірності: жодне просте технологічне рішення не є вільним від непередбачуваних побічних ефектів; боротьба з дезінформацією має системні обмеження; а розуміння поведінки користувачів в алгоритмічно керованому середовищі є недостатнім.

Скандал, пов'язаний із діяльністю компанії Cambridge Analytica у березні 2018 року – незаконне отримання даних понад 87 мільйонів користувачів Facebook для психографічного профілювання виборців [8] – спричинив перехід соцмереж і Facebook в першу чергу до більш структурованих підходів боротьби із фейками. Програма партнерства для перевірки фактів розширилася до понад 80 партнерів у 60 країнах [14]. Одночасно із цим запроваджене алгоритмічне зниження видимості ідентифікованих фейків зменшило їхнє охоплення в середньому на 80% [1]. Проте структурна проблема цієї боротьби залишилася незмінною: фактчекери не змогли охопити весь проблемний контент, а найбільше вірусне охоплення та поширення фейк отримував у перші хвилини після публікації [12].

Twitter та YouTube обрали різні стратегії щодо боротьби з фейками. Twitter (нині X) перейшов від видалення контенту до його пояснення: запровадив маркування маніпулятивних матеріалів (2020) і систему Community Notes – контекстуальні виправлення чи спростування, які з'являються лише після досягнення консенсусу між користувачами з різними поглядами [28]. YouTube зазнав критики через ефект «кролячої нори» в рекомендаціях – тенденцію алгоритму послідовно пропону-

вати дедалі радикальніший контент – і у 2019 році знизив пріоритетність сумнівного контенту: за даними компанії, перегляди таких матеріалів через рекомендації в США скоротилися на 70% [27]. Водночас базова мета алгоритму – максимізація часу перегляду – залишилася незмінною.

Пандемія COVID-19 породила так звану інфодемію – лавиноподібне поширення неперевіреної та шкідливої інформації про вірус, лікування і вакцини, що становило пряму загрозу громадському здоров'ю. Це змусило платформи вперше запровадити проактивне видалення фейкового контенту: Facebook видалив понад 20 млн матеріалів і позначив як сумнівні 190 млн публікацій [11], YouTube заблокувала більше 1 млн відео [5], Twitter видалив понад 97 тис. публікацій щодо дезінформації про вакцини [30]. Водночас агресивна модерація призвела до блокування легітимного контенту, зокрема, наукових дискусій про походження коронавірусу, які платформи спочатку кваліфікували як конспірологію, а згодом відмовились від цієї ідеї [16].

Вибори 2020 року в США поглибили фундаментальну дилему модерації: якщо після 2016 року платформи критикували за бездіяльність, то в 2020-му – за сильніше втручання. Превентивні заходи з боку платформ включали виборчі інформаційні центри [19], заборону нової політичної реклами [20], попередні мітки щодо підозр про фальсифікації [13]. Після 6 січня 2021 року Twitter і Facebook заблокували акаунт чинного президента, спровокувавши гострі дебати про межі між модерацією та цензурою.

Протягом 2018–2022 років платформи усвідомили міжплатформенний характер інформаційних операцій і налагодили неформальну співпрацю через посередників – Atlantic Council DFRLab, Stanford Internet Observatory, Graphika [9]. Звіти про прозорість [29] зафіксували еволюцію операцій: від очевидних ботів, що поширювали фейки у 2016–2017 роках до складних мереж акаунтів у 2020–2022-му [22].

До кінця 2022 року існуючі підходи до виявлення фейків досягли структурних меж. Ручна модерація не могла масштабуватися до мільярдних аудиторій [23], алгоритми ефективно виявляли спам і фейкові акаунти, але залишалися ненадійними щодо напівправди, вирваного з контексту матеріалу та конспірологічних наративів. Паралельно технології deepfake та ШІ-інструменти продемонстрували здатність генерувати дезінформаційний контент у промислових масштабах [7], а Digital Services Act ЄС 2022 року створив регуляторне підґрунтя для інвестицій у ШІ-інструменти модерації. Таким чином, до кінця досліджуваного періоду боротьба з дезінформацією вступила в принципово новий етап [1].

Висновки. Проведений аналіз дозволяє зробити кілька ключових висновків щодо природи та динаміки боротьби з фейковим контентом у соціальних мережах протягом 2015–2022 років.

По-перше, досліджуваний період засвідчив перехід від реактивних тактичних рішень до системних підходів, однак цей перехід відбувався переважно під примусом – скандалів, регуляторного тиску та суспільної критики – й не був результатом проактивної позиції платформ. Кожен новий інструмент – від позначок фактчекерів до алгоритмічної депріоритизації – мав непередбачені побічні ефекти й вимагав постійного коригування. Це свідчить не лише про складність проблеми, а й про відсутність у платформ комплексного розуміння природи дезінформації як системного явища.

По-друге, бізнес-модель соціальних мереж, орієнтована на максимізацію залученості, структурно суперечить завданням протидії дезінформації: емоційно насичений і маніпулятивний контент органічно переважає в алгоритмічних рекомендаціях. Без зміни фундаментальних стимулів у бізнес-моделях платформ будь-які технічні інструменти модерації матимуть обмежену ефективність.

По-третє, до кінця 2022 року існуючі підходи досягли структурних меж: масштаби дезінформаційних загроз систематично перевищували спроможність реагування як платформ, так і фактчекерської екосистеми. Поява загальнодоступних інструментів генеративного ШІ унеможливила подальше покладання на рішення, розроблені для принципово іншого технологічного середовища, і створила передумови для якісно нового етапу в боротьбі з дезінформацією.

По-четверте, досліджуваний період засвідчив якісну еволюцію самих дезінформаційних загроз: від очевидних автоматизованих мереж ботів 2016–2017 років до складних мереж скоординованої неавтентичної поведінки 2020–2022 років, які було важче виявити. Ця еволюція вимагала від платформ переходу від видалення окремих матеріалів до пошуку мереж акаунтів, що зумовило розвиток міжплатформенної співпраці.

Перспективи подальших досліджень охоплюють кілька пріоритетних напрямів: аналіз ефективності ШІ-орієнтованих підходів до модерації контенту в умовах генеративного ШІ після 2022 року

та у нових регуляторних умовах, зокрема, в контексті DSA ЄС; специфіку дезінформаційних операцій в україномовному просторі та адаптацію глобальних інструментів протидії до вітчизняного контексту; а також роль медіаграмотності як довгострокового чинника суспільної стійкості до дезінформаційних впливів у порівнянні з технологічними та регуляторними підходами.

Список літератури:

1. Баловсяк Н. Більше модераторів і власне відео: як Facebook продовжує боротися з фейками. *StopFake*. URL: https://www.stopfake.org/uk/bilshe-moderatoriv-i-vlasne-video-yak-facebook-prodovzhuye-borotysya-z-fejkamy/
2. Почепцов Г. Г. Сучасні інформаційні війни. Київ : Видавничий дім «Києво-Могилянська академія», 2015. 495 с.
3. Світова гібридна війна: український фронт : монографія / за заг. ред. В. П. Горбуліна. Харків : Фоліо, 2017. 496 с.
4. Allcott H., Gentzkow M. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*. 2017. Vol. 31, No. 2. P. 211–236. DOI: https://doi.org/10.1257/jep.31.2.211.
5. Bell K. YouTube has removed 1 million videos for dangerous COVID-19 misinformation. *TechCrunch*. 2021. 25 August. URL: https://techcrunch.com/2021/08/25/youtube-has-removed-1-million-videos-for-dangerous-covid-19-misinformation/
6. Bradshaw S., Howard P. N. *The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation*. Oxford : Oxford Internet Institute, University of Oxford, 2019. URL: https://comprop.oii.ox.ac.uk/research/cybertroops2019/
7. Buchanan B., Lohn A., Musser M., Sedova K. *Truth, Lies, and Automation: How Language Models Could Change Disinformation*. Washington, D.C. : Center for Security and Emerging Technology (CSET), Georgetown University, 2021. DOI: https://doi.org/10.51593/2021CA003.
8. Cadwalladr C., Graham-Harrison E. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica. *The Guardian*. 2018. 17 March. URL: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
9. DiResta R. et al. *The Tactics & Tropes of the Internet Research Agency*. New Knowledge, 2018. URL: https://int.nyt.com/data/documenthelper/533-read-report-internet-research-agency/7871ea6d5b7bedafbf19/optimized/full.pdf
10. EUvsDisinfo (European External Action Service). Annual reviews of disinformation attribution operations. 2019–2022. URL: https://euvsdisinfo.eu/
11. Facebook. An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19. *Meta Newsroom*. 2021. URL: https://about.fb.com/news/2020/04/covid-19-misinfo-update/
12. Funke D. In the past year, Facebook has quadrupled its fact-checking partners. *Poynter*. 2019. 25 April. URL: https://www.poynter.org/fact-checking/2019/in-the-past-year-facebook-has-quadrupled-its-fact-checking-partners/
13. Google LLC (YouTube). Supporting the 2020 U.S. election. *YouTube Blog*. 2020. 9 December. URL: https://blog.youtube/news-and-events/supporting-the-2020-us-election/
14. Hard Questions: How Is Facebook's Fact-Checking Program Working? *Meta Newsroom*. 2018. 21 June. URL: https://about.fb.com/news/2018/06/hard-questions-fact-checking/
15. Isaac M. Facebook, after criticism, drops 'disputed' label for fake news. *The New York Times*. 2017. 20 December. URL: https://www.nytimes.com/2017/12/20/business/media/facebook-fake-news.html
16. Krawchuk C. et al. Research note: Examining how various social media platforms have responded to COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*. 2022. URL: https://misinforeview.hks.harvard.edu/article/research-note-examining-how-various-social-media-platforms-have-responded-to-covid-19-misinformation/
17. Lyons T. Replacing disputed flags with related articles. *Facebook Newsroom*. 2017. 20 December. URL: https://techcrunch.com/2017/12/20/facebook-will-ditch-disputed-flags-on-fake-news-and-display-links-to-trustworthy-articles-instead/
18. Pennycook G., Bear A., Collins E. T., Rand D. G. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*. 2020. Vol. 66, No. 11. P. 4944–4957. DOI: https://doi.org/10.1287/mnsc.2019.3478.
19. Rosen G. Preparing for Election Day. *Meta Newsroom*. 2020. 7 October. URL: https://about.fb.com/news/2020/09/preparing-for-election-day/
20. Rosen G. Update on our election integrity efforts. *Meta Newsroom*. 2020. 21 October. URL: https://about.fb.com/news/2020/10/update-on-our-election-integrity-efforts/

21. Shearer E., Gottfried J. News Use Across Social Media Platforms 2017. *Pew Research Center*. 2017. URL: https://www.pewresearch.org/journalism/2017/09/07/news-use-across-social-media-platforms-2017/

22. Stanford Internet Observatory. *Coordinated Inauthentic Behavior Reports*. 2020–2022. URL: https://stacks.stanford.edu

23. Statista. *Number of monthly active users of leading social networks worldwide as of 2022*. Statista, 2022. URL: https://statista.com

24. Stretch C. Testimony of Colin Stretch, General Counsel, Facebook : hearing before the U.S. Senate Select Committee on Intelligence. Washington, D.C., 2017. 1 November. URL: https://www.judiciary.senate.gov/imo/media/doc/10-31-17%20Stretch%20Testimony.pdf

25. Struhárik F. Biggest drop in Facebook organic reach we have ever seen. *Medium*. 2017. 21 October. URL: https://medium.com/@filip_struharik/biggest-drop-in-organic-reach-weve-ever-seen-b2239323413

26. Tandoc E. C., Lim Z. W., Ling R. Defining «Fake News»: A typology of scholarly definitions. *Digital Journalism*. 2018. Vol. 6, No. 2. P. 137–153. DOI: https://doi.org/10.1080/21670811.2017.1360143.

27. The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation. *YouTube Blog*. 2019. 3 December. URL: https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/

28. Twitter (X Corp.). Building rules in public: Our approach to synthetic and manipulated media. 2020. URL: https://blog.x.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media

29. Twitter Safety. *Information Operations on Twitter* : серія звітів прозорості. 2018–2022. URL: https://transparency.twitter.com

30. Twitter. Updates to our work on COVID-19 vaccine misinformation. *Twitter Blog*. 2021. URL: https://blog.x.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation

31. Vosoughi S., Roy D., Aral S. The spread of true and false news online. *Science*. 2018. Vol. 359, No. 6380. P. 1146–1151. DOI: https://doi.org/10.1126/science.aap9559.

32. Wardle C., Derakhshan H. *Information disorder: Toward an interdisciplinary framework for research and policy making*. Strasbourg : Council of Europe, 2017. 109 p. URL: https://rm.coe.int/information-disorder-report-207/1680766412

Balovsiak N. V. FIGHTING FAKE CONTENT ON SOCIAL NETWORKS IN 2015–2022: A HISTORICAL ASPECT

The article provides a comprehensive historical analysis of social media approaches to countering fake content from 2015 to 2022, before the emergence and widespread use of generative artificial intelligence technology. The study covers the transformation of strategies to combat disinformation from chaotic experiments by individual platforms to relatively systematic, but still fragmented, moderation mechanisms. The key stages of social media policy formation are analyzed: the reaction to the 2016 US presidential election, the 2018 Cambridge Analytica scandal, the COVID-19 pandemic, and the events surrounding the 2020 US elections. The main tools for countering disinformation are systematized: manual moderation, partnerships with fact-checkers, algorithmic de- prioritization of problematic content, labeling of manipulative materials, and detection of coordinated information operations. The evolution of each of these tools over time, their advantages, and structural limitations in the face of growing volumes of user-generated content are separately considered. The specifics of the approaches of the most popular platforms to combat fake news – Facebook, Twitter and YouTube – are examined and a comparative analysis of their chosen moderation strategies is carried out, taking into account differences in business models, technical infrastructure and the nature of the audience.

Particular attention is paid to the systemic limitations of the chosen models of countering disinformation, in particular the conflict between the commercial interests of the platforms and the public demand for reliable information, as well as the problem of uneven application of moderation rules depending on the geographical region, language of the content and media significance of the actors. The issues of transparency and accountability of platforms to society and regulators are investigated, as well as discussions around the limits of permissible interference by private companies in the public information space. It has been proven that the reactive nature and insufficient scalability of the approaches of that time, as well as the emergence of publicly available generative artificial intelligence (AI) tools after 2022, created the prerequisites for the transition to automated AI-oriented solutions in the field of disinformation detection and neutralization.

Keywords: *fake content, disinformation, social networks, content moderation, fact-checking, information operations, artificial intelligence, hybrid threats.*

Дата першого надходження статті до видання: 16.04.2026
Дата прийняття статті до друку після рецензування: 01.05.2026
Дата публікації (оприлюднення) статті: 30.05.2026