Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка

Факультет інформаційних технологій
Кафедра кібербезпеки та захисту інформації

ДОПУСТИТИ ДО ЗАХИСТУ:
В.о. завідувача кафедри
кібербезпеки
та захисту інформації
_____ Іван ПАРХОМЕНКО
« » _____ 2025 р.

ПОЯСНЮВАЛЬНА ЗАПИСКА

кваліфікаційної роботи

галузь знань *12 Інформаційні технології*
(шифр і назва галузі знань)

спеціальність *125 Кібербезпека та захист інформації*
(код і назва спеціальності)

освітній ступень *магістр*

освітньо-наукова програма *Кібербезпека*
(назва освітньої програми)

на тему: «Модель впливу комплексної протидії технології Deepfake»

Виконавець: студент ІІ курсу, групи КБм-21

(підпис) **Максим ПРИЩЕПА** (Ім’я, ПРІЗВИЩЕ)

	Ім’я, ПРІЗВИЩЕ	**Підпис**
Науковий керівник	Володимир НАКОНЕЧНИЙ	
Нормоконтроль	Іван БІЛОКОНЬ	

Київ 2025

Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка

Факультет інформаційних технологій
Кафедра кібербезпеки та захисту інформації

ЗАТВЕРДЖЕНО:
В.о. завідувача кафедри
кібербезпеки
та захисту інформації
_____ Іван ПАРХОМЕНКО
«25» жовтня 2024 р.

ЗАВДАННЯ
на виконання кваліфікаційної роботи

спеціальності *125 Кібербезпека та захист інформації*
(код і назва спеціальності)

освітній ступень *магістр*

Здобувача(ки) КБм-21 (група) Прищепи Максима Олександровича (прізвище ім'я по-батькові)

Тема кваліфікаційної роботи «Модель впливу комплексної протидії технології Deepfake»

1. ПІДСТАВИ ДЛЯ ПРОВЕДЕННЯ РОБОТИ

Рішення засідання кафедри кібербезпеки та захисту інформації факультету інформаційних технологій протокол № 4 від 24.10.2024 р.

2. МЕТА ТА ВИХІДНІ ДАНІ ДЛЯ ПРОВЕДЕННЯ РОБІТ

Об'єкт досліджень	Процес ведення кібервійн.
Предмет досліджень	Використання Deepfake як методу ведення кібервійн та модель впливу комплексної протидії технології Deepfake.
Мета	Дослідження технології Deepfake як методу ведення сучасних кібервійн.
Вихідні дані для проведення роботи	Методи ведення сучасних кібервійн та методи протидії технології Deepfake.

3. ОЧІКУВАНІ НАУКОВІ РЕЗУЛЬТАТИ

Наукова новизна	Дослідження місця та ролі технології Deepfake у сучасних кібервійнах та проведення моделювання з використанням нечіткої логіки у два етапи: модель впливу Deepfake на інформаційну безпеку держави та впливу комплексної протидії технології Deepfake.
Практична цінність	Виконана модель має дозволяти проводити сценарне моделювання для подальшого вивчення впливу Deepfake та впливу протидії, а також для покращення методів протидії Deepfake.

4. ЕТАПИ ВИКОНАННЯ РОБОТИ

Найменування етапів робіт	Строки виконання робіт (початок-кінець)
Уточнення постановки задачі	25.10.2024 – 29.12.2024
Аналіз літературних джерел	30.12.2024 – 09.03.2025
Ознайомлення з сучасними реаліями та тенденціями розвитку та використання Deepfake	10.03.2025 – 16.03.2025
Дослідження наявних матеріалів про Deepfake, його походження та передумови виникнення, й існуючі технології генерації	17.03.2025 – 23.03.2025
Дослідження застосування Deepfake в якості одного з методів ведення кібервійн	24.03.2025 – 30.03.2025
Розробка моделі впливу Deepfake на інформаційну безпеку держави та моделі впливу комплексної протидії технології Deepfake	31.03.2025 – 13.04.2025
Проведення сценарного моделювання за розробленою моделлю та аналіз отриманих результатів	14.04.2025 – 27.04.2025
Оформлення пояснювальної записки згідно методичних рекомендацій	28.04.2024 – 15.05.2025
Подача пакету документів на розгляд ЕК	15.05.2025 – 19.05.2025

Завдання видав _____ Володимир НАКОНЕЧНИЙ
(підпис) (Ім'я, ПРІЗВИЩЕ)

Завдання прийняв
до виконання _____ Максим ПРИЩЕПА
(підпис) (Ім'я, ПРІЗВИЩЕ)

Дата видачі завдання: 25.10.2024 р.
Термін подання кваліфікаційної роботи до ЕК 19.05.2025 р.

РЕФЕРАТ

Пояснювальна записка до кваліфікаційної роботи «Модель впливу комплексної протидії технології Deepfake» має обсяг 80 сторінок, містить 31 рисунок, 4 таблиці та 44 джерела.

Об’єкт дослідження: процес ведення кібервійн.

Предмет дослідження: використання технології Deepfake як методу ведення кібервійн, а також модель впливу комплексної протидії технології Deepfake.

Мета дослідження: дослідження технології Deepfake як методу ведення сучасних кібервійн.

Для досягнення зазначеної мети дипломної роботи поставлені такі завдання дослідження:

- дослідження Deepfake, його походження та передумови виникнення, існуючі технології генерації та застосування в якості одного з методів ведення кібервійн;
- розробка моделі впливу Deepfake на інформаційну безпеку держави та моделі впливу комплексної протидії технології Deepfake;
- проведення сценарного моделювання за розробленою моделлю та аналіз отриманих результатів.

Методи дослідження: методи аналізу та синтезу, порівняння, моделювання, класифікації та кластеризації.

Наукова новизна та практичне значення: у роботі досліджено місце та роль технологій Deepfake у сучасних кібервійнах та проведено моделювання з використанням нечіткої логіки у два етапи: модель впливу Deepfake на інформаційну безпеку держави та впливу комплексної протидії технології Deepfake. Виконана модель дозволяє проводити сценарне моделювання для

подальшого вивчення впливу Deepfake та впливу протидії, а також для покращення методів протидії Deepfake.

Перспектива подальших досліджень порушеної у цій роботі проблематики полягає у глибшому аналізі впливу технології Deepfake та впливу протидії, що дозволить моделювати та прогнозувати можливі сценарії та впливи – це є надзвичайно важливим для ефективної та своєчасної протидії.

Ключові слова: кібервійна, кіберзброя, інформаційна війна, deepfake, діпфейк, нейронні мережі, інформаційно-психологічний вплив, дезінформація, нечітка логіка, нечітка когнітивна карта, fuzzy cognitive map.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

APT – Advanced Persistent Threat
(D)DoS – (Distributed) Denial-of-Service
GAN – Generative adversarial network
NATO – North Atlantic Treaty Organization
НКК – Нечітка когнітивна карта
РНБО – Рада національної безпеки і оборони [України]
(Ш)ПЗ – (Шкідливе) програмне забезпечення

ЗМІСТ

ВСТУП

Зі стрімким розвитком цифрових технологій та глобалізацією інформаційного простору кордони між державами у віртуальному світі стають дедалі більш умовними. Соціальні мережі, месенджери та відеохостинги давно перетворилися на головні майданчики для поширення інформації, де власниками акаунтів та каналів можуть бути як офіційні медіа, так і невідомі особи, зокрема учасники ворожих інформаційних операцій. У таких умовах маніпулятивний контент може легко проникати в масову свідомість, оскільки його первинна перевірка зазвичай відбувається вже після публікації. Це створює ідеальне середовище для інформаційних та інформаційно-психологічних атак, які стають ключовим елементом сучасних кібервійн.

Останні досягнення у сфері штучного інтелекту, зокрема розвиток глибинного навчання (англ. deep learning), суттєво розширили арсенал засобів інформаційного або інформаційно-психологічного впливу. Одним із найнебезпечніших інструментів, що набув широкого розповсюдження, є технологія Deepfake. Використовуючи великі масиви даних для навчання, нейронні мережі здатні створювати фальшиві відео та зображення, які виглядають вражаюче реалістично. На відміну від традиційного редагування в графічних редакторах, алгоритми Deepfake можуть точно відтворювати міміку, голос і рухи людини, підміняючи її особистість у цифровому середовищі.

У контексті кібервійни Deepfake може бути застосований для створення фальшивих звернень лідерів держав, генерації підроблених доказів, поширення паніки серед населення та інших застосувань. Небезпека цієї технології полягає в тому, що, поєднуючи Deepfake із іншими методами ведення кібервійн – фішингом, соціальною інженерією, бот-мережами тощо – можна створювати складні багаторівневі інформаційні та інформаційно-психологічні атаки. У результаті суспільство стає вразливішим до дезінформації, що може загрожувати не лише окремим індивідам, а й національній безпеці в цілому.

Для формування теоретичного підґрунтя було проаналізовано відповідну літературу й публікації за темою. Зокрема, публікації Центру передового досвіду в галузі стратегічних комунікацій НАТО (англ. NATO Strategic Communications Centre of Excellence) «Deepfakes – Primer and Forecast»[1] та «The Role of Deepfakes in Malign Influence Campaings»[2], «Кримінологічний аналіз використання технології Deepfake: коли фейк стає злочином» Юртаєвої К.В.[3], «Діпфейк та дезінформація» Вальорскої М.А.[4].

Для розуміння місця технологій Deepfake в сучасних кібервійнах було розглянуто та опрацьовано огляд Бурака Джинара «Deepfakes in Cyber Warfare: Threats, Detection, Techniques and Countermeasures» [5], статтю науковців McGill University «Deep Fakes and Big Data: the Next Level of Cyber Warfare» [6], статтю науковців University College Cork «Deepfakes in warfare: new concerns emerge from their use around the Russian invasion of Ukraine» [7], а також комплексний огляд українського досвіду в кібервійні «A Decade in the Trenches of Cyberwarfare: Ukraine's Story of Resilience» [8].

Більшість дослідників, які публікують пов'язані з Deepfake наукові матеріали, фокусуються на таких аспектах: технології виявлення підробок Deepfake, дезінформаційні наслідки підробок Deepfake та юридичні аспекти законного врегулювання застосування Deepfake. Наразі вкрай мало наукових матеріалів, де автори розглядають та досліджують Deepfake як нову небезпечну складову сучасних кібервійн. Адресуючи цю прогалину, запропоновано дослідити Deepfake як метод ведення кібервійн, а також змоделювати вплив комплексної протидії технології Deepfake, яка має на меті протистояння впливу Deepfake на інформаційну безпеку держави.

Об'єкт дослідження: процес ведення кібервійн.

Предмет дослідження: використання технології Deepfake як методу ведення кібервійн, а також модель впливу комплексної протидії технології Deepfake.

Мета дослідження: дослідження технології Deepfake як методу ведення сучасних кібервійн.

Для досягнення зазначеної мети дипломної роботи поставлені такі завдання дослідження:

- дослідження Deepfake, його походження та передумови виникнення, існуючі технології генерації та застосування в якості одного з методів ведення кібервійн;
- розробка моделі впливу Deepfake на інформаційну безпеку держави та моделі впливу комплексної протидії технології Deepfake;
- проведення сценарного моделювання за розробленою моделлю та аналіз отриманих результатів.

Методи дослідження: методи аналізу та синтезу, порівняння, моделювання, класифікації та кластеризації.

Наукова новизна та практичне значення: у роботі досліджено місце та роль технологій Deepfake у сучасних кібервійнах та проведено моделювання з використанням нечіткої логіки у два етапи: модель впливу Deepfake на інформаційну безпеку держави та впливу комплексної протидії технології Deepfake. Виконана модель дозволяє проводити сценарне моделювання для подальшого вивчення впливу Deepfake та впливу протидії, а також для покращення методів протидії Deepfake.

Перспектива подальших досліджень порушеної у цій роботі проблематики полягає у глибшому аналізі впливу технології Deepfake та впливу протидії, що дозволить моделювати та прогнозувати можливі сценарії та впливи, що є надзвичайно важливим для ефективної та своєчасної протидії.

Апробація результатів дослідження: Прищепа М. О., Наконечний В. С. Модель впливу комплексної протидії технології Deepfake. Проблеми кібербезпеки інформаційно-комунікаційних систем: Збірник матеріалів доповідей та тез; м. Київ, 11 квітня 2025 року; Київський національний університет імені Тараса Шевченка / Редкол.: В.В. Ільченко, д.ф-м.н., проф., (голова); та ін. – К.: ВПЦ "Київський університет", 2025. – с. 10–12.

РОЗДІЛ 1
ПОТОЧНИЙ СТАН ПРОБЛЕМАТИКИ DEEPFAKE

1.1 Походження, термінологія та інформаційно-психологічний вплив

Термін «Deepfake» утворено шляхом поєднання двох англомовних термінів: «deep learning» (глибинне навчання) та «fake» (несправжній, підробка), що дослівно означає «підробка, створена за допомогою глибинного навчання». Під цим поняттям розуміють синтетичні аудіо-, фото- та відеофайли, згенеровані за допомогою нейронних мереж, які здатні з високою точністю імітувати реальні об’єкти чи події. Цікаво, що саме під цим псевдонімом у 2017 році діяв анонімний користувач платформи «Reddit», який став відомим завдяки поширенню порнографічних відеопідробок, у яких замінював обличчя реальних людей на зображення знаменитостей [9]. Серед тих, хто став жертвами його експериментів, опинилися відомі акторки, такі як Скарлетт Йоганссон, Галь Гадот і Тейлор Свіфт [10].

Використовуючи алгоритми глибинного навчання, він підміняв обличчя у відео, використовуючи зображення, знайдені у відкритому доступі в Інтернеті. Якість таких підробок напряму залежала від обсягу матеріалів, доступних для тренування нейромережі. Зрештою адміністрація платформи заблокувала його акаунт, але технологія вже привернула увагу широкої аудиторії. Програми для створення Deepfake швидко поширилися в мережі, викликавши серйозні дискусії щодо їхнього правового регулювання. Спочатку ця технологія використовувалася переважно для розваг, але згодом вона стала інструментом маніпуляцій, шантажу та навіть кібератак.

Сучасний розвиток алгоритмів глибинного навчання зробив Deepfake доступним для широкого кола користувачів. Якщо раніше для створення реалістичних відеопідробок потрібні були навички професійних фахівців і значні ресурси, то сьогодні це можна зробити за допомогою спеціального програмного

забезпечення, використовуючи лише персональний комп'ютер. Це суттєво підвищує ризики застосування таких технологій у деструктивних цілях, особливо в контексті інформаційних війн, кібервійн та комбінованих кібератак.

Крім того, варто звернути увагу на потенціал використання технології Deepfake для здійснення цілеспрямованого інформаційно-психологічного впливу як на окремих осіб, так і на великі соціальні групи. Такий вплив складається з двох ключових компонентів: інформаційного та психологічного. У своїй статті «Інформаційне насильство, інформаційна маніпуляція та пропаганда: поняття, ознаки та співвідношення» дослідники та Фурашев В.М. та Самчинська О.А. визначають інформаційно-психологічний вплив як поєднання цих двох складових. Отже, дослідники вказують: «Інформаційно-психологічний вплив – це цілеспрямований, переважно організований процес проникнення у свідомість людини (або групи осіб), який здійснюється за допомогою сукупного використання спеціальних інформаційних засобів та технологій і психологічних прийомів та спрямований на зміну індивідуальних та/або групових психічних явищ та (або) психічний або фізичний стан людини (або групи осіб)» [11]. Вони наголошують, що це цілеспрямований вплив на природний хід психічних процесів людини, що базується на використанні інформації, інформаційних технологій, методів обробки даних, а також як вербальних, так і невербальних, а також паралінгвістичних та інших психологічних засобів. Такий вплив зазвичай спрямовується на свідомість окремої особи або суспільства в цілому, реалізується за допомогою інформаційно-психологічних чи інших методів, і веде до змін у психіці: трансформації світогляду, уявлень, відношення до подій, системи цінностей, мотивації та усталених стереотипів. Основна мета такого впливу – викликати конкретну поведінкову реакцію, що узгоджується з цілями впливу.

Інформаційно-психологічний вплив відрізняється від суто психологічного тим, що він завжди організований і поряд із психологічними засобами передбачає використання спеціальних інформаційних технологій. Від інформаційного впливу він відмежовується тим, що, крім передачі певних даних, спрямований безпосередньо на зміну психічного стану людини або групи осіб. Водночас такий

вплив може не тільки викликати зміни у свідомості чи емоційному стані, а й опосередковано позначитися на інформаційній інфраструктурі певного об’єкта.

Структура інформаційно-психологічного впливу складається з кількох основних компонентів:

- суб’єкта (хто ініціює вплив);
- об’єкта (на кого спрямований);
- мети (чого прагне ініціатор);
- методів і засобів (як здійснюється вплив);
- кінцевого результату.

Роль суб’єкта можуть виконувати різні актори – як окремі особи чи групи, так і державні органи, міжнародні організації, спецслужби або приватні структури. Головне, що їхні дії завжди спрямовані на досягнення певних цілей за допомогою конкретних інструментів.

Щодо об’єкта впливу, то головною мішенню завжди є люди – як окремі особи, так і цілі соціальні групи. Однак існують і непрямі об’єкти впливу, такі як суспільні настрої, соціальні процеси, ставлення до подій чи явищ, система цінностей, а також функціонування державних інституцій та бізнесу. Оскільки інформаційно-психологічний вплив є не випадковим, а завчасно підготовленим процесом, його ключовим елементом є чітке визначення мети. Важливу роль також відіграють методи та засоби реалізації цього впливу, що можуть охоплювати широкий спектр психологічних і технологічних прийомів, серед яких, зокрема, займають своє місце технології створення реалістичних фото- та відеопідробок Deepfake.

1.2 Deepfake як кіберзброя

Ще одним важливим питанням є визначення того, чи можна вважати Deepfake формою кіберзброї (англ. cyberweapon). На сьогодні існує кілька

підходів до тлумачення цього терміну, що зумовлено різними поглядами дослідників на характер і спосіб застосування таких технологій.

Перший підхід розглядає кіберзброю виключно з технічної точки зору. У цьому випадку під нею розуміється сукупність технічних та програмних засобів, які використовуються для експлуатації вразливостей інформаційних і телекомунікаційних систем противника.

Другий підхід фокусується на соціальному аспекті, трактуючи кіберзброю як набір технологій, призначених для впливу на суспільство або окремі групи людей через цифрові платформи.

Третій, узагальнений підхід, поєднує обидва попередні, визначаючи кіберзброю як комплекс засобів, здатних здійснювати вплив як на кібернетичні системи, так і на суспільні процеси через кіберпростір. У якості прикладу третього підходу наведемо визначення кіберзброї, що приведене авторами підручника «Основи кібербезпеки і кібероборони». Автори надають таке визначення: «Кіберзброя – це набір технічних, програмних та інших засобів, спрямованих на порушення процесів управління в кіберпросторі, включаючи соціум, соціотехнічні системи, технічні системи (комп'ютерні системи та мережі, системи зв'язку та автоматизовані системи управління, управляючі елементи систем озброєння і військової техніки та небезпечних об'єктів і об'єктів з критичною інформаційною інфраструктурою, програмне забезпечення, бази даних тощо) у вигляді інформаційних, психологічних та різноманітних фізичних деструктивних впливів» [12].

З урахуванням комплексного третього підходу до визначення поняття «кіберзброя», технологію Deepfake можна розглядати як кіберзброю, яка має технічне походження. Вона належить до програмних засобів і передбачає здійснення впливу, наприклад інформаційно-психологічного, на окремих індивідів, групи людей або суспільство загалом.

Деякі дослідники наголошують на тому, що соціальні аспекти відіграють ключову роль у застосуванні кіберзброї. Кібервійна охоплює не лише боротьбу за контроль над інформаційними потоками, а й цілеспрямоване формування

громадської думки, маніпулювання суспільними настроями та підрив довіри до державних і міжнародних інституцій.

Терористичні угруповання та окремі держави дедалі частіше використовують соціальні медіа як ефективний інструмент у веденні кібервійн [13]. Ці платформи дозволяють швидко поширювати дезінформацію, впливати на емоційний стан аудиторії, координувати деструктивні дії та навіть вербувати нових прихильників. У цьому контексті Deepfake стає особливо небезпечним інструментом, оскільки дає змогу створювати реалістичні фальшиві відео, що можуть змінювати хід суспільних дискусій або дискредитувати окремих осіб та організації.

Крім того, у своєму дослідженні «The Dark Side of Interconnectivity: Social Media as a Cyber–Weapon?», опублікованому в книзі «Social Media and the Armed Forces», дослідниця Софія Мартінс дійшла висновку, що соціальні медіа можуть використовуватися як кіберзброя. Більше того, Російська Федерація вже застосовувала такий підхід проти України станом на 2020 рік. А саме через платформи соціальних мереж в Інтернеті здійснювалася розвідка, вербування найманців до складу незаконних збройних формувань на тимчасово окупованих територіях Донецької та Луганської областей, а також проводилися інформаційні та психологічні операції на різних рівнях, у тому числі на політичному до військовому рівнях [14].

1.3 Методи генерації підробок Deepfake

Спосіб створення Deepfake може багато розповісти про його джерело. Як правило, для генерації таких підробок застосовують два основних підходи.

Першим таким підходом є використання генеративно-змагальних нейронних мереж (англ. generative adversarial network – GAN), концепція яких була розроблена дослідниками Університету Монреаля у 2014 році [15]. Ці мережі мають широкий спектр можливостей, оскільки їх можна навчити генерувати майже будь-який тип даних. Після навчання нейронна мережа здатна

самостійно створювати нові дані – зображення, аудіо, відео або навіть багатовимірні об’єкти – на основі отриманого навчального набору, фактично виконуючи функцію генератора.

Основний принцип роботи GAN полягає у взаємодії генератора і дискримінатора – це дві нейромережі, які функціонують у форматі змагання. Генератор створює штучні дані, а дискримінатор оцінює їхню автентичність, порівнюючи зі справжніми прикладами. Процес навчання ґрунтується на двосторонньому зворотному зв’язку: генератор поступово вдосконалюється, виробляючи все реалістичніші підробки, а дискримінатор, у свою чергу, підвищує точність у їх розпізнаванні. На рисунку 1.1 представлено загальну схему роботи генеративно-змагальної нейронної мережі.

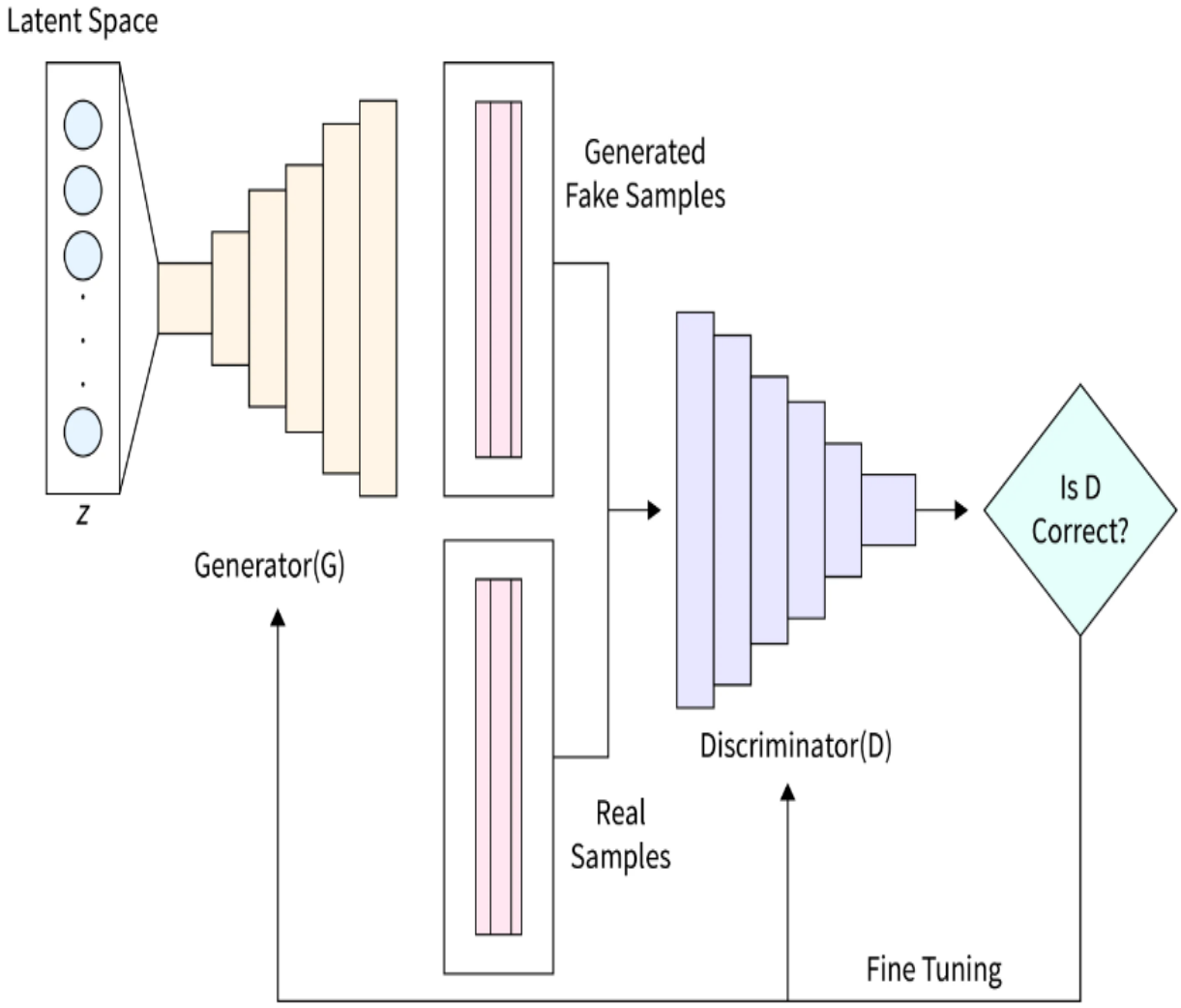

Рисунок 1.1 – Схема GAN

Ще одним популярним інструментом на основі нейронних мереж, який активно використовується для створення підробленого контенту, є автокодер (англ. autoencoder). Зазвичай він реалізується у вигляді двох пар «кодер-декодер» (англ. encoder-decoder), де обидві кодуючі частини мають однакову структуру.

Історично автокодери розроблялися як моделі для виявлення складних нелінійних залежностей між елементами даних і мали здатність відтворювати зображення, на яких вони були попередньо навчені.

На вході кодер стискає дані зображення, кодуючи його спеціальні атрибути, такі як текстура й колір шкіри, вираз обличчя, стан очей (відкриті чи закриті), положення голови та інші дрібні особливості. Отримане стиснене представлення зображення передається до прихованого простору (англ. latent space), який дозволяє моделі вивчати внутрішні закономірності та структурні подібності між різними точами даних. На завершальному етапі декодер розкодовує цю інформацію, намагаючись відновити оригінальне зображення на основі його стисненої репрезентації.

Отже, декодер прагне реконструювати зображення таким чином, щоб воно максимально відповідало оригіналу. Така архітектура автокодера створює так званий «ефект пляшкового горла» (англ. bottleneck), коли інформація примусово стискається до найважливіших ознак. У результаті на виході кодера формується компактне представлення даних, відоме як «карта особливостей» (англ. feature map), а в контексті Deepfake — як «приховане обличчя» (англ. latent face).

У результаті такого стиснення на карті характеристик зберігається лише найважливіша інформація, необхідна для якісної реконструкції, тоді як другорядні або зайві дані відкидаються. Загальний принцип роботи автокодера зображено на рисунку 1.2.

Щоб використати архітектуру автокодера для створення Deepfake-підробок, необхідно застосувати дві пари кодер-декодер з однаковими кодуючими частинами. Така конструкція дає змогу кодеру виявляти спільні риси між двома наборами зображень. Під час навчання кожна пара тренується окремо.

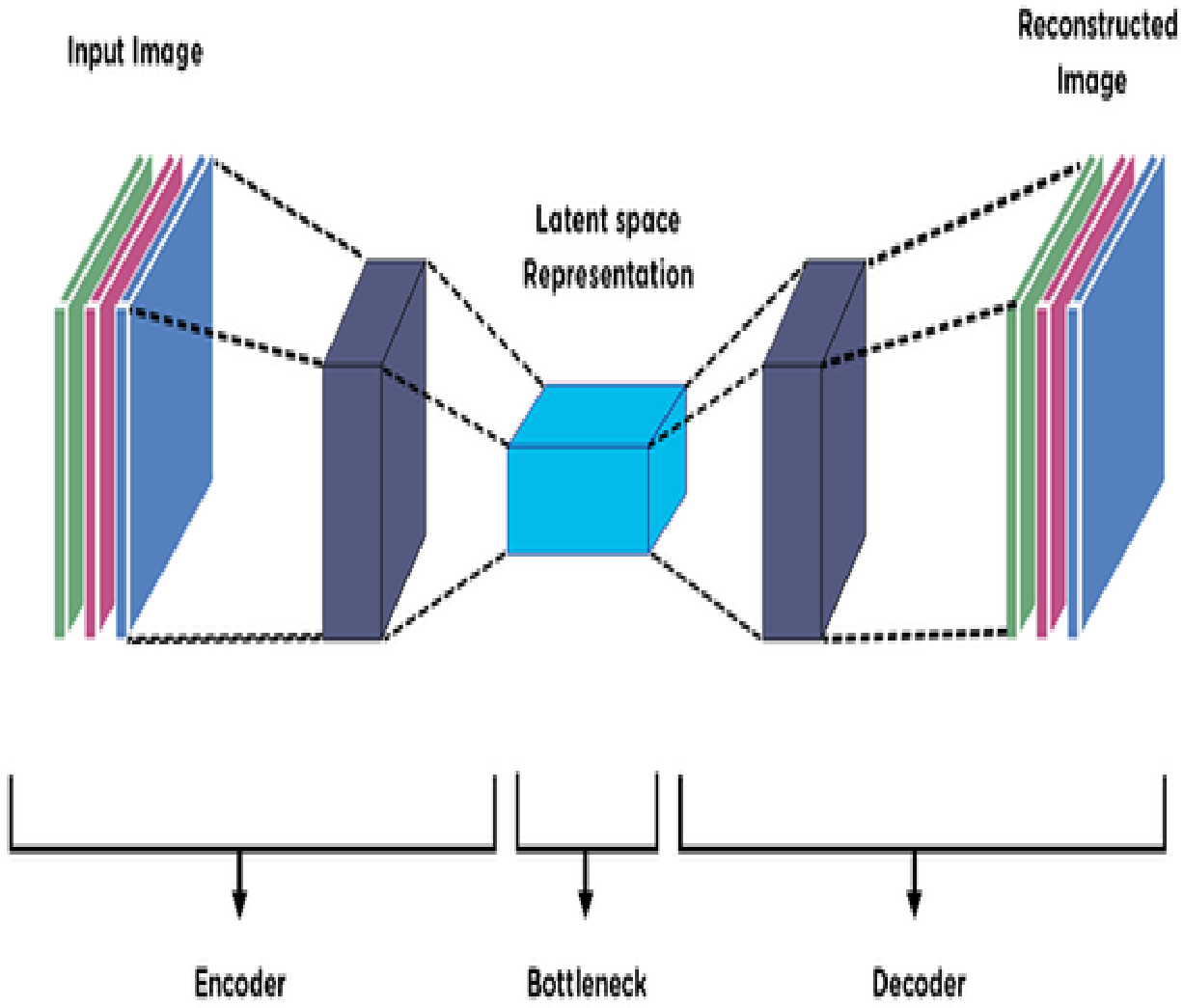

Рисунок 1.2 – Схема Autoencoder

На завершальному етапі декодувальні частини мережі змінюють місцями, що дозволяє генерувати зображення одного класу, використовуючи карти особливостей, отримані з іншого класу зображень [16]. На рисунку 1.3 представлено загальну схему процесу створення Deepfake-підробок з використанням архітектури Autoencoder.

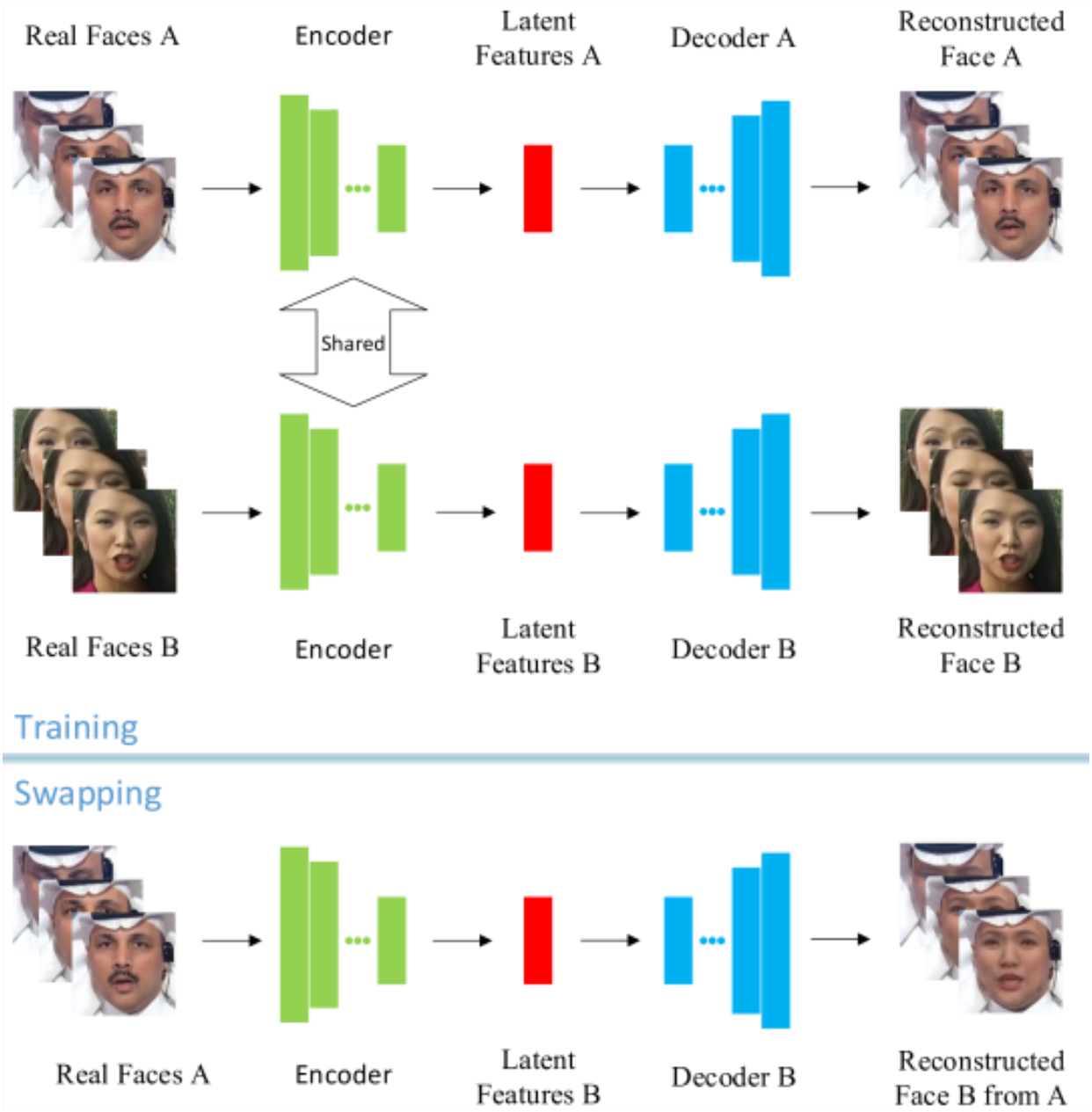

Рисунок 1.3 – Метод генерації Deepfake-підробки за допомогою автокодера

Таким чином, маємо підтвердження, що технологія Deepfake є кіберзброєю, що має технічне походження, оскільки її можна віднести до програмних засобів. Це зумовлено тим, що для створення таких підробок Deepfake використовуються сучасні технології на основі нейронних мереж, зокрема генеративно-змагальні нейронні мережі (GAN) та автокодери, що можуть забезпечувати високий рівень реалістичності підробленого контенту.

Висновки до розділу 1

У першому розділі проаналізовано сучасний стан проблематики технологій Deepfake . Огляд теоретичних понять і термінології, пов’язаних із діпфейками, а також дослідження двох основних методів генерації підробок Deepfake показали, що ці технології можна класифікувати як кіберзброю, здатну чинити інформаційно-психологічний вплив на окремих осіб, групи осіб чи цілі суспільства. Завдяки розвитку технологій глибинного навчання, зокрема генеративно-змагальних нейронних мереж і автокодерів, створення переконливих підробок стало значно простішим, не вимагаючи високих навичок та значного досвіду редагування чи потужних обчислюваньних ресурсів. Поширення підробок Deepfake у соціальних медіа, разом із значним прогресом у їх породжує серйозні виклики, які вимагають комплексних рішень для ефективної протидії.

РОЗДІЛ 2
DEEPFAKE ЯК МЕТОД ВЕДЕННЯ КІБЕРВІЙН

2.1 Основні поняття та проблематика кібервійн

У сучасному світі інформаційні технології стали невід'ємною частиною життя, проте водночас вони відкрили новий фронт для конфліктів – кіберпростір. На відміну від традиційних воєн, кібервійни ведуться не на полі бою, а у віртуальному середовищі, де основною зброєю є цифрові технології, інформаційні маніпуляції та кібератаки. Держави, терористичні організації, хакерські угруповання та навіть окремі індивіди використовують кіберінструменти для досягнення своїх політичних, економічних і військових цілей.

Методи ведення кібервійн включають широкий спектр тактик і стратегій, які можна умовно поділити на технічні та інформаційно-психологічні. До технічних методів належать, до прикладу, кібератаки на критичну інфраструктуру, системи зв'язку, фінансові установи, державні ресурси тощо. Вони можуть проявлятися у вигляді DDoS-атак, впровадження шкідливого програмного забезпечення, використання вразливостей у програмних продуктах та зламу інформаційних систем. Інформаційно-психологічні методи, своєю чергою, спрямовані на вплив на громадську думку, маніпуляцію масовою свідомістю та поширення дезінформації. Одним із найнебезпечніших інструментів у цій сфері стало використання технологій нейронних мереж штучного інтелекту, зокрема Deepfake, для створення фальшивих зображень, відео та аудіоматеріалів, які складно відрізнити від реальних.

Сучасні кібервійни поєднують ці методи, створюючи новий тип загроз, які важко ідентифікувати та ще складніше нейтралізувати. Вони можуть бути спрямовані як на державні інститути, так і на окремих громадян, створюючи нестабільність у суспільстві, підриваючи довіру до офіційних джерел інформації

та впливаючи на політичні процеси. У такому середовищі розуміння методів ведення кібервійн є ключовим для розробки ефективних стратегій кібербезпеки та захисту національних інтересів.

Одразу зазначимо, що англомовні автори розрізняють поняття «Cyberwar» та «Cyberwarfare». Різниця криється у значенні слів в англійській мові, хоча на українську мову обидва терміни перекладаються як кібервійна. В англійській мові слово «War» позначає сам військовий конфлікт між певною кількістю сторін, а «Warfare» – діяльність, пов'язану з веденням війни, часто включаючи зброю та методи, які використовуються [17].

Більш того, термін «кібервійна» часто викликає суперечки серед військових експертів, і багато хто з них уникає його вживання, замінюючи більш конкретними поняттями. Наприклад, у «Концептуальному плані розвитку можливостей сухопутних військ з ведення кібероперацій в кіберпросторі на період з 2016 по 2028 роки», який був опублікований у 2010 році Командуванням навчально-наукового центру з розбудови сухопутних військ США, замість терміну «кібервійна» використовується поняття «бойова операція в кіберпросторі» (англ. Cyber Warfare Operation) [18]. Такий підхід свідчить про прагнення військових структур чіткіше визначати характер і межі застосування кіберінструментів у військових конфліктах. Це також відображає складність конфліктів у кіберпросторі, у яких межа між різними діями та методами часто є розмитою.

Адресуючи вищенаведені проблеми термінології, українські дослідники у підручнику «Основи кібербезпеки і кібероборони» вводять два поняття «кібервійна» та «кібероперація». Наведемо ці визначення: «Кібероперація – це скоординована й узгоджена за масштабом, місцем і часом паралельна або послідовна кібердія розвідувального, оборонного та (або) наступального характеру, яка має на меті завоювання переваги в кіберпросторі за рахунок нанесення збитків суб’єктам та об’єктам з критичною інформаційною інфраструктурою протиборчої сторони та захисту власних кібернетичних систем від аналогічних дій у відповідь» [19]. «Кібервійна – це складне суспільно-

політичне явище, що протікає у вигляді конфлікту в кіберпросторі між протиборчими сторонами (державами, коаліціями держав тощо) з використанням кібернетичних систем; високотехнологічний конфлікт, продовження політики держав і (або) коаліцій, політичних угруповань, транснаціональних корпорацій і т.д. з метою нав'язати опонентам свою волю за допомогою впливу на них у кіберпросторі і через нього в різних сферах життєдіяльності у формі кіберпротиборства, військових (кібер-) дій між їх кіберсилами (кібервійськами) [19]». У найближчій перспективі кібервійна може стати окремим і самостійним елементом широкомасштабних воєнних дій, поряд із застосуванням наземних, повітряних, морських і космічних сил [12]. Її характерними рисами є складність або неможливість встановлення джерела агресії; прихований характер впливу та відсутність явних фізичних руйнувань на початкових етапах; надзвичайно висока швидкість здійснення атак, коли час між початком дій і їхніми наслідками мінімізується. Сучасна кіберзброя не залежить від географічних меж чи відстаней, а також практично не зустрічає технологічних, правових та інших бар'єрів. Крім того, їй притаманні синергетичні, ланцюгові, вторинні, третинні та подібні ефекти, що значно ускладнюють оцінку масштабу і наслідків таких дій.

Дослідник Kenneth Boyte у своєму матеріалі «The Evolution of Cyber Warfare in Information Operations Targeting Estonia, the U.S., and Ukraine» для книги «Developments in Information Security and Cybernetic Wars» досліджував три кейси ведення кіберовійн (cyberwarfare) [20]. Кібервійна стала невід’ємною частиною сучасних конфліктів, інтегруючись у гібридну війну, що поєднує кінетичні та некінетичні методи. Boyte підкреслює, що ведення cyberwarfare в Естонії, США та Україні демонструють здатність впливати на критичну інфраструктуру (енергомережі, комунікації), політичні процеси (вибори) та суспільну думку, створюючи хаос і дезорієнтацію.

Більш, того атаки в Естонії, США та Україні характеризуються анонімністю виконавців і складністю визначення відповідальності. Хоча підозри вказують на Росію (Естонія, Україна) та Іран (США), брак публічних доказів ускладнює збір доказової бази та звинувачення сторони нападу. Це забезпечує

"правдоподібне заперечення" для державних і недержавних акторів, розширюючи поле для кібероперацій.

У цьому матеріалі чітко простежується тенденція еволюції cyberwarfare – кібервійна еволюціонувала від простих хакерських атак до складних багаторівневих операцій, що загрожують критичній інфраструктурі, політичній стабільності та людському життю. Наведені автором кейси Естонії, США та України ілюструють зростання технологічної складності, анонімності та впливу кібератак, що вимагає комплексних підходів до кібероборони.

2.2 Методи ведення кібервійн

Описуючі суто технічні аспекти сучасних кібервійн та проявів Cyberwarfare фахівці компанії «Avast» виділяють наступні об’єкти атак [21]:

- Цивільна інфраструктура, наприклад, електромережі або системи управління дорожнім рухом;
- Фінансові установи, такі як банки, інвестиційні фонди та кредитні спілки;
- Військові об'єкти, а також підрядники та інші установи, діяльність яких пов’язана із забезпеченням національної безпеки;
- Окремі громадяни цільової країни.

Серед вагомих наслідків технічних зловмисних дій в рамках Cyberwarfare розрізняють [21]:

- Збої в електропостачанні – порушення роботи національної електромережі може завдати шкоди економіці та вплинути на громадську думку;
- Порушення кібербезпеки – хакерські атаки можуть пошкодити програмні системи або скомпрометувати чутливі урядові мережі;
- Витоки даних – масштабні витоки даних можуть вплинути на цілий ряд персональних даних, таких як медичні записи або банківські реквізити;

- Військовий або промисловий саботаж – прямі атаки на національну безпеку або економічну інфраструктуру країни погіршують військовий або промисловий потенціал;
- Порушення зв'язку – телефонний, мобільний, електронний або інший цифровий зв'язок може бути вимкнений, перехоплений або іншим чином порушений;

Розглянемо основні актуальні виклики, які виділяють дослідники при дослідженні кібервійн та Cyberwarfare [22]:

1. Системи раннього попередження – складність полягає у визначенні того, на що має бути спрямована система раннього попередження у кіберпросторі. Перед розробниками таких систем постають наступні питання, на які система має дати відповіді при оповіщенні:
   - Чи відбувається кібервійна вже зараз або ось-ось розпочнеться?
   - Яка тактика атаки застосовується?
   - Хто є агресором?
   - Яка передбачувана мета супротивника?
2. Закони та етика кібервійни: кібервійна, як і будь-яка інша дія, що може завдати шкоди, створює етичні проблеми та виклики. Зокрема, держави повинні розуміти, чи є кібервійна етично прийнятною і як вести таку війну етично – у відповідності до законів війни;
3. Ведення кібервійни: коли з'являється новий вимір для ведення війни, визначення того, як на ній ефективно функціонувати, стає першочерговим завданням. Дослідники пропонують вісім нових принципів ведення кібернетичної війн:
   - Подвійне використання методів ведення кібервійн;
   - Мутабельність і непослідовність атак;
   - Кіберпростір як оперативне середовище;
   - Кінетичні ефекти дій у кіберпросторі;
   - Ідентифікація привілеїв як ключ для реалізації атак;

- Здійснення контролю над інфраструктурою;
- Відсутність фізичних обмежень;
- Прихованість дій;

4. Кіберзброя: поняття кіберзброї викликає цілий ряд питань, зокрема, як визначити кіберзброю чи можливо регулювати її розробку і застосування. Наразі не існує такого єдиного відходу, який би був застосовний у всьому світі.

5. Проблеми атрибуції: атрибуція визначається як «визначення особи або місцезнаходження зловмисника або посередника зловмисника». На думку авторів, атрибуція є критично важливим компонентом кібервійни, оскільки багато наступальних методів ведення кібервійн можуть бути застосовані у відповідь на атаку у кіберпросторі лише в тому випадку, якщо джерело атаки можна встановити з високою достовірністю.

6. Стримування у кіберпросторі: постає питання, як можна перешкодити агресору розпочинати зловмисні дії в кіберпросторі? Це питання, як і питання атрибуції, стосується всіх видів кібератак загалом, а не лише кібервійни. Оскільки агресор зазвичай невідомий, традиційні правила важко застосувати.

7. Погляди центральних органів влади країн на кібервійну: розуміння підходів, що застосовуються урядами країн як головними організаторами війн, є важливим завданням. Літературні джерела чітко показують, що країни усвідомлюють проблему кібервійн і намагаються розробити власні шляхи протидії та здійснення дій у кіберпросторі.

8. Концептуалізація кібервійни – цей виклик є дещо ширшим за вищезгадані виклики і має на меті представити кілька точок зору на кібервійну. Деякі дослідники пропонують концепцію кібервійни як гри, в якій може взяти участь кожен, хто має пристрій, підключений до Інтернету. На їх думку, будь-хто може брати участь у кібервійні, але найбільший вплив все-таки мають держави.

Отже, на основі досліджених джерел про природу кібервійн та Cyberwarfare складемо перелік методів, які використовуються на даний час для ведення кібервійн та Cyberwarfare. Розрізнятимемо три ключові групи методів, які принципово відрізняються один від одного за цілями та природою дій. Методи кібервійн мають ділитися на:

1. Наступальні операції – активні цілеспрямовані дії на саботаж, нанесення шкоди, ведення підривну діяльність, перешкоджання або зрив діяльності противника [22];

2. Операції з розвідки – отримання інформації з обмеженим доступом (засекреченої, службової або конфіденційної інформації) корпорацій, урядів, приватних осіб, організацій тощо з метою отримання військової, політичної або економічної вигоди шляхом використання злочинних методів експлуатації комп'ютерів, програмного забезпечення, мереж та Інтернету, фішингу тощо;

3. Операції впливу – це стратегічні дії, спрямовані на зміну поведінки, переконань або рішень цільової аудиторії шляхом маніпулювання інформацією. Вони можуть включати розповсюдження пропаганди, дезінформацію, використання бот-мереж та алгоритмів соціальних платформ для підсилення певних наративів, а також інформаційно-психологічний вплив на окремих осіб або суспільство в цілому. Такі операції нерідко супроводжуються атаками на довіру до офіційних джерел інформації, дискредитацією державних інституцій чи громадських діячів.

Для проведення наступальних операцій користуються наступними методами:

- Атаки за допомогою шкідливого програмного забезпечення: Це включає використання вірусів, черв’яків, троянів та інших видів шкідливого ПЗ для пошкодження чи порушення роботи систем. Шкідливе ПЗ може бути використано для непоправної шкоди життєво важливим комп'ютерним та телекомунікаційним системам [23].

- Атаки типу "відмова в обслуговуванні" (DoS і DDoS): Ці атаки спрямовані на перевантаження систем, щоб зробити їх недоступними. DDoS-атаки часто використовуються для блокування доступу до високопрофільних веб-серверів, таких як банки.
- Фішинг і соціальна інженерія: Ці методи передбачають обман користувачів для отримання доступу до систем чи викрадення даних. Фішинг може бути використаний для компрометації чутливих комп'ютерних систем [24].
- Атаки на критичну інфраструктуру: Ці атаки спрямовані на життєво важливі системи, такі як електромережі, водопостачання чи транспорт. Міжнародний комітет Червоного Хреста підкреслює ризик шкоди цивільним від кібероперацій, які впливають на критичну інфраструктуру, як-от телекомунікації чи фінансові системи [25].
- Атаки з використанням «вразливостей нульового дня» (англ. zero-day exploits): Це експлуатація раніше невідомих вразливостей у програмному забезпеченні для отримання доступу чи контролю. Atlantic Council згадує, що 5 вразлвостей нульового дня було використано в ізраїльській атаці Stuxnet на іранську атомну промисловість [26].
- Атаки на ланцюги постачання: Компрометація програмного забезпечення чи апаратного забезпечення до його доставки цільовому об'єкту. Такі атаки можуть дозволити зловмисникам отримати доступ після розгортання продукту [21].
- Внутрішні (інсайдерські) загрози (англ. insider threats): Використання чи примус працівників для отримання доступу чи завдання шкоди. Ворожі уряди та інші потенційні зловмисники можуть використовувати незадоволених чи недбалих працівників [27].
- Розвинені стійкі загрози (англ. APTs – advanced persistent threats): Це довгострокові, цільові атаки, часто приписувані державам (хакерським угрупуванням, яких пов'язують з державами), які включають комбінацію технік

для підтримки доступу до системи. Наприклад, APT-операції під час російсько-української війни [28].

Операції з розвідки зазвичай використовують один метод, який описують загальним терміном, хоча конкретний набір тактичних прийомів та технік може значно відрізнятися:

- Кібершпигунство (англ. espionage) – крадіжка чутливої інформації для стратегічної чи економічної переваги. Кібератаки, які саботують урядові комп'ютерні системи, можуть бути використані для підтримки звичайних військових дій. Такі атаки можуть блокувати офіційний урядовий зв'язок, заражати цифрові системи, уможливлювати крадіжку життєво важливих розвідувальних даних і загрожувати національній безпеці. Наприклад, державні або військові атаки можуть бути спрямовані на військові бази даних, щоб отримати інформацію про розташування військ, озброєння і техніку, що використовується [24].

Операції впливу це відносно нова група методів ведення кібервійн та Cyberwarfare, яка ще вивчається дослідниками та вченими. Згрупуємо відомі методи операцій впливу у три ключові методи:

- Дезінформаційні та пропагандистські кампанії – організовані поширення маніпулятивної або повністю неправдивої інформації для досягнення політичних, військових чи економічних цілей. Такі кампанії можуть спрямовуватися на створення хаосу, підрив довіри до урядів, інституцій чи демократичних процесів, а також на формування вигідного наративу серед цільової аудиторії [24].
- Кампанії психологічного та інформаційно-психологічного впливу – цілеспрямовані дії, що використовують психологічні та інформаційні методи для формування, зміни або підриву переконань, емоційного стану та поведінки цільової аудиторії. Такі кампанії можуть мати на меті деморалізацію населення чи військових, створення соціальної напруженості, посилення розколу в суспільстві або стимулювання вигідних для ініціатора рішень і дій.

- Маніпуляції соціальними мережами та бот-мережами – систематичне використання автоматизованих акаунтів (ботів), фейкових профілів, популярних інфлюенсерів та алгоритмів соціальних платформ для розповсюдження маніпулятивної інформації, формування певних наративів, впливу на громадську думку та створення ілюзії широкої підтримки певних ідей та/або наративів [29].

2.3 Місце технологій Deepfake серед методів ведення кібервійн

На питання, чи є підробки Deepfake прикладами методів ведення кібервійн можна можемо висунути припущення, що Deepfake є складовими методів ведення кібервійн, перш за все методів операцій впливу. Дослідимо наукові джерела, автори яких досліджували Deepfake у контексті кібервійн.

На користь на користь висунутого вище припущення наводять тези науковці McGill University «Deep Fakes and Big Data: the Next Level of Cyber Warfare» [6]. Автори розглядають Deepfake як метод кібервійн, зокрема в контексті інформаційних операцій, що впливають на національну безпеку та демократичні процеси. У документі прямо згадується, що deepfake використовуються в "спонсорованих державами дезінформаційних кампаніях" (англ. state-sponsored disinformation campaigns), які є частиною кібервійн. Зазначається, що ці кампанії сприяють поляризації суспільства, підривають демократію та можуть провокувати радикальні дії в цільових групах, таких як діаспори чи екстремістські рухи. Низька вартість і легкий доступ до технології роблять deepfake привабливим інструментом для акторів кібервійн. Вони посилюють інформаційні операції, які є одним з ключових елементів сучасних кібервійн, які спрямовані на маніпуляцію сприйняттям і дестабілізацію суспільства.

Крім того, Бурак Джинар у своєму огляду «Deepfakes in Cyber Warfare: Threats, Detection, Techniques and Countermeasures» [5] досліджує загрози

deepfake для суспільства. Автор розглядає Deepfake як інструмент кіберзлочинності та пропаганди, що має прямий зв'язок із кібервійнами, особливо в контексті операцій впливу та дезінформації. Документ пов'язує Deepfake із кіберзлочинністю та інформаційними операціями, які є складовими кібервійн. Наприклад, кіберзлочинці використовують Deepfake для імперсонування (наприклад, генеральних директорів), щоб обдурити працівників на переказ грошей. Також автор наводить приклад пропагандистської операції «Spamouflage» (про-китайська кампанія), яка використовувала боти з Deepfake для поширення фальшивого контенту в соціальних мережах, що є прикладом інформаційної війни, яка часто є частиною кібервійн.

Схожий висновок надають і науковці University College Cork у статті «Deepfakes in warfare: new concerns emerge from their use around the Russian invasion of Ukraine» [7] – автори слушно зазначають, що природа підробок Deepfake робить цю технологію такою, що добре підходить для використання під час ведення кібервійн (англ. cyberwarfare). Вчені розглядали підробки Deepfake, які з'явилися у 2022 році після початку повномасштабної фази війни. Один з головних їх висновків, які вони змогли сформувати, спостерігаючи за підробками Deepfake в соціальних медіа – надмірна кількість підробок Deepfake в соцмережах підривають довіру до всіх типів медіа, оскільки справжні відео потім користувачі можуть помилково називатися Deepfake-підробками. Більш того, низькотехнологічні фейки, як-от відео з хибними субтитрами чи кадри з інших війн, додатково ускладнюють інформаційний простір. Автори акцентують увагу на участь технологій Deepfake у поєднанні з наступальними операціями у кіберпросторі, таких як хакерські атаки на популярні ЗМІ – таким чином довірене джерело після зламу зловмисниками може транслювати підробку Deepfake з метою впливу на аудиторію джерела. Насамкінець, науковці стверджують, що існуючі технології виявлення Deepfake наразі не є самодостатнім рішенням для боротьби з підробками, не останню чергу через те, що поширюються підробки Deepfake набагато швидше, аніж спрацьовує моніторинг, перевірка відео на достовірність та інформування про підробку. На

їх думку, для боротьби з впливом Deepfake під час кібервійн важливо заохочувати належну медіаграмотність людей, щоб користувачі балансували між здоровим і нездоровим скептицизмом. У результаті користувачі мають скептично ставитися до надто провокаційних матеріалів і чекати, поки такі відомості будуть підтверджені кількома джерелами, яким можна довіряти. З іншого боку, люди повинні бути обережними, щоб не звинувачувати будь-яке відео в тому, що воно є підробкою, та не втрачати довіри до окремих медіа через засилля підробок Deepfake в інформаційному просторі.

Отже, на основі аналізу наукових джерел маємо підтвердження, що Deepfake, безперечно, є методом кібервійн (cyberwarfare), особливо він посідає важливе місце в операціях впливу.

Висновки до розділу 2

У другому розділі технології Deepfake були розглянуті у якості складових сучасних кібервійн. Зокрема, було досліджено сучасний науковий підхід до термінології та опису кібервійн, у тому числі розділення англомовними дослідниками термінів cyberwar та cyberwarfare, що позначають стан (кібервійни) та дії (під час кібервійни) відповідно. Досліджено та описано сучасні методи ведення кібервійн, що органічно об’єднуються у три групи: наступальні операції, операції розвідки та операції впливу – які принципово відрізняються одна від одної за цілями. Після чого, обґрунтовано приналежність технологій Deepfake до групи методів операцій впливу, оскільки частіше за все технології Deepfake використовуються для дезінформації, що має на меті введення в оману цільової групи, та для інформаційно-психологічних впливів, які націлені на зміну поведінки та/або переконань цільової групи.

РОЗДІЛ 3
МОДЕЛЮВАННЯ ВПЛИВУ КОМПЛЕКСНОЇ ПРОТИДІЇ DEEPFAKE

1.1 Модель впливу Deepfake на інформаційну безпеку

Моделювання та аналіз ведення кібервійн та інформаційно-психологічного впливу використовуваної кіберзброї є складним завданням через, те що така система включає слабоструктуровані ситуації. Розв'язання цієї задачі можливе завдяки застосуванню когнітивного підходу, який дозволяє уникнути жорсткої формалізації та використовувати експертні судження для формування гнучких моделей. Суть когнітивного підходу полягає у створенні моделей, що відображають знання про ситуацію у вигляді причинно-наслідкових взаємозв'язків між ключовими факторами. Одним із прикладів таких моделей є нечіткі когнітивні карти, що фактично виступають математичною інтерпретацією досліджуваної проблеми. Ідея побудови нечіткої когнітивної карти (англ. fuzzy cognitive map) була вперше запропонована Бартом Коско [30]. Така карта являє собою орієнтований граф, де вершини відповідають концептам (факторам), орієнтовані дуги – причинно-наслідковим зв'язкам між цими концептами, а ваги дуг характеризують ступінь впливу одного концепту на інший. У загальному вигляді нечітку когнітивну карту можна описати наступною формулою:

$$Fuzzy\ Cognitive\ Map = < C, F, W > \qquad (3.1)$$

де C – множина концептів карти (англ. concepts), F (англ. fuzzy relations) – множина дуг карти, W – множина всіх ваг (англ. weights).

Формальна когнітивна карта дещо відрізняється, оскільки дуги такої карти можуть приймати фіксовані значення позитивного та негативного впливу. Наведемо подання такої карти:

$$Cognitive\ map = < C, F >, \quad (3.2)$$

Якщо у формальній когнітивній карті існує відношення A_{ij} між вершинами C_i та C_j, то у випадку якщо C_i та C_j, пов'язані один з одним – A_{ij} приймає значення 1, якщо вершини не пов'язані один з одним – A_{ij} приймає значення 0. Таке відношення A_{ij} може приймати значення «+1» або «–1» – це залежить від направленості дуги, яка описує відношення між вершинами C_i та C_j.

Для побудови нечіткої когнітивної карти, яка відображатиме вплив технології Deepfake на інформаційну безпеку держави, передусім необхідно визначити множину ключових концептів. У процесі аналізу та дослідження різних аспектів Deepfake було сформовано набір найбільш значущих концептів, перелік котрих представлено в таблиці 3.1.

Таблиця 3.1 – Обрані концепти нечіткої когнітивної карти

Коротке позначення	Концепт
C1	Сприймана достовірність підробки
C2	Недовіра до діючої політичної влади
C3	Недовіра до офіційних джерел інформації
C4	Якісна підробка Deepfake
C5	Прийняття інформації через відеоконтент
C6	Довіра до джерела отримання підробки
C7	Перевірка відео на достовірність в інших джерелах
C8	Інформаційно-психологічний вплив на цільову групу осіб
C9	Неприйняття змісту підробки
C10	Політична криза та (або) зміна вектору державної політики діючої влади
C11	Ефект праймінгу (priming effect)

Наступним кроком є визначення ваг дуг у межах інтервалу $W_{ij} \in [-1;1]$. Для встановлення причинно-наслідкових відношень між концептами використовується спеціально розроблена шкала, яка дозволяє оцінити як характер зв'язку (позитивний чи негативний), так і його силу. Оцінювання виконується за допомогою лінгвістичної шкали, де кожному її значенню відповідає певний числовий діапазон: для позитивних впливів – у межах [0; 1], а для негативних – у межах [–1; 0]. Нижче подано множину лінгвістичних значень, що використовуються для оцінки ваги впливу одного концепту на інший :

$$\textit{Вага впливу} = \{\textit{Не впливає; Дуже слабка; Слабка; Середня; Сильна; Дуже сильна}\} \quad (3.3)$$

Далі кожному з лінгвістичних значень слід зіставити певний числовий інтервал. У результаті формується схема перетворення якісної оцінки, наданої експертом, у нечітке кількісне значення. Відповідність між цими оцінками подано в таблиці 3.2 нижче.

Таблиця 3.2 – Оцінка ваг впливу між концептами

Лінгвістична змінна	Діапазон значень
Негативний дуже сильний вплив	[-1; -0,85)
Негативний сильний вплив	[-0,85; -0,6)
Негативний середній вплив	[-0,6; -0,35)
Негативний слабкий вплив	[-0,35; -0,15)
Негативний дуже слабкий вплив	[-0,15; 0)
Вплив між концептами відсутній	0
Позитивний дуже слабкий вплив	(0; 0,15]
Позитивний слабкий вплив	(0,15; 0,35]
Позитивний середній вплив	(0,35; 0,6]
Позитивний сильний вплив	(0,6; 0,85]
Позитивний дуже сильний вплив	(0,85; 1]

Оскільки вже визначено множину концептів та здійснено градацію ваг причинно-наслідкових зв’язків між ними, можна перейти до побудови графічної частини нечіткої когнітивної карти, використавши застосунок MentalModeler [31]. На Рисунку 3.1 представлено результат моделювання, що включає множину концептів та відповідні орієнтовані дуги графа, які відображають взаємозв’язки між ними.

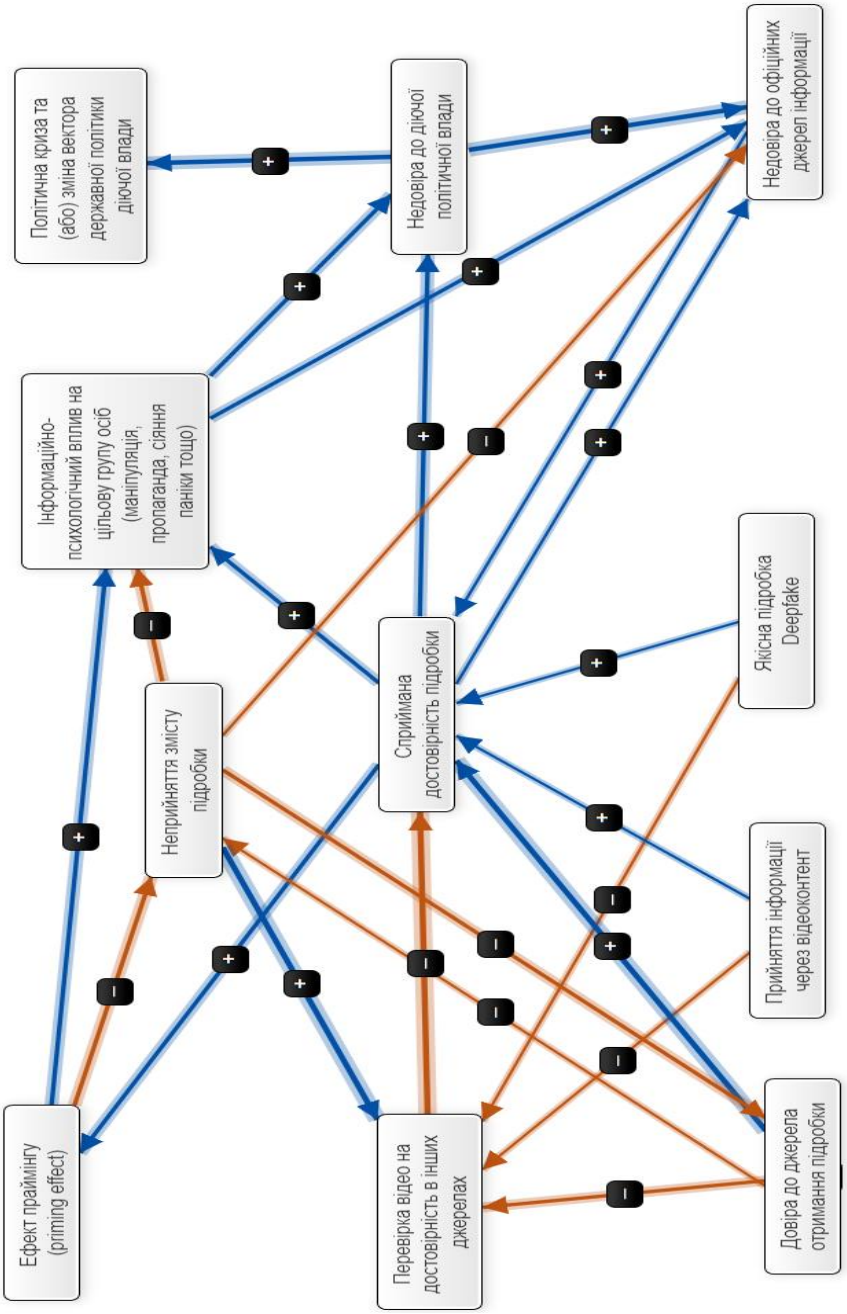

Рисунок 3.1 – Нечітка когнітивна карта для дослідження впливу Deepfake на інформаційну безпеку держави

З графом моделі асоціюється матриця суміжності W, яка наведена у формулі 3.4. Кожен елемент *Wij*, розташований на перетині *i*-го рядка та *j*-го стовпця цієї матриці, відображає характер та ступінь впливу концепту C_i на концепт C_j. На основі цього сформуємо матрицю суміжності, яка представляє взаємозв'язки між концептами у межах побудованої нечіткої когнітивної карти.

$$W = \begin{vmatrix} W_{11} & \dots & W_{1n} \\ \dots & \dots & \dots \\ W_{n1} & \dots & W_{nn} \end{vmatrix} \quad (3.4)$$

Визначення вагових коефіцієнтів зв'язків між концептами здійснюється на основі експертного оцінювання, що проводиться спеціальною експертною комісією. Це дозволяє встановити значення W_{ij} для кожної пари концептів. Спочатку рівень впливу визначається експертно шляхом використання лінгвістичних оцінок із відповідної множини (див. формулу 3.3).

Для забезпечення точності аналізу, виконання обчислень основних параметрів моделі та здійснення подальших операцій з нею, лінгвістичні оцінки сили впливу потребують переведення у відповідні числові значення.

Згідно з обраним підходом, експертна комісія узгоджує числові значення ваг взаємовпливів між концептами на основі попередньо визначених лінгвістичних оцінок, орієнтуючись на відповідності, подані у Таблиці 3.2.

Використовуючи такий підхід, залучена комісія, що складалася зі старших аналітиків компанії ТОВ «Умбра. Дослідження та аналітика», погодила числові значення вагових коефіцієнтів зв'язків між концептами. Умовні позначення цих значень наведено в Таблиці 3.3, де використовуються позначення вигляду C_i, де *i* набуває цілих значень від 1 до 11 – відповідно до порядку концептів (див. Таблицю 3.1), а матриця суміжності з повними назвами концептів подана на Рисунку 3.2.

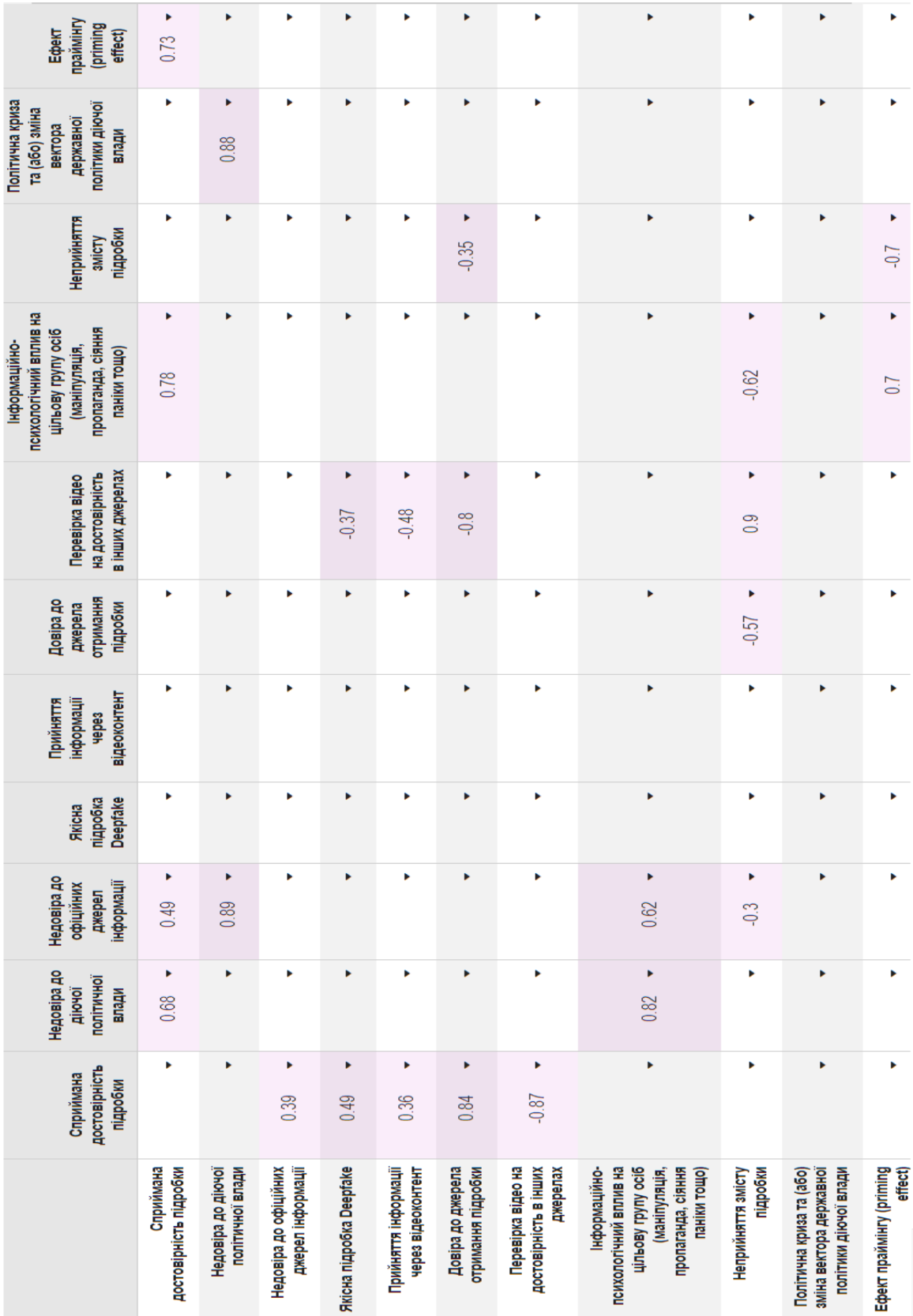

Рисунок 3.2 – Матриця взаємовпливів концептів нечіткої когнітивної карти

Таблиця 3.3 – Матриця взаємовпливів концептів нечіткої когнітивної карти у скороченому записі

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
C1	0	0,68	0,49	0	0	0	0	0,78	0	0	0,73
C2	0	0	0,89	0	0	0	0	0	0	0,88	0
C3	0,39	0	0	0	0	0	0	0	0	0	0
C4	0,49	0	0	0	0	0	-0,37	0	0	0	0
C5	0,36	0	0	0	0	0	-0,48	0	0	0	0
C6	0,84	0	0	0	0	0	-0,80	0	-0,35	0	0
C7	-0,87	0	0	0	0	0	0	0	0	0	0
C8	0	0,82	0,62	0	0	0	0	0	0	0	0
C9	0	0	-0,30	0	0	-0,57	0,90	-0,62	0	0	0
C10	0	0	0	0	0	0	0	0	0	0	0
C11	0	0	0	0	0	0	0	0,70	-0,70	0	0

Обчислимо коефіцієнт кластеризації для розробленої моделі. Він дає можливість визначити зв’язність та оцінити щільність зв’язків розробленої нечітної когнітивної крати. Для обчислення коефіцієнту кластеризації маємо використати формулу 3.5:

$$D = \frac{M}{N(N-1)} \tag{3.5}$$

де M – кількість всіх зв’язків між концептами карти, а N – кількість вершин.

У нашому випадку кількість концептів N=11, а кількість всіх зв’язків між вершинами карти M=23.

Підставляючи ці значення у формулу 3.5, розраховуємо значення коефіцієнта кластеризації: $D = \frac{23}{110} = 0{,}2091$.

Отримане значення коефіцієнта кластеризації свідчить про достатній рівень зв'язності побудованої моделі. На Рисунку 3.3 представлено ключові показники даної нечіткої когнітивної карти, яка моделює вплив технології Deepfake на інформаційної безпеки держави.

Component	Indegree	Outdegree	Centrality
Сприймана достовірність підробки	2.95	2.6799999999999997	5.63
Недовіра до діючої політичної влади	1.5	1.77	3.27
Недовіра до офіційних джерел інформації	2.3	0.39	2.69
Якісна підробка Deepfake	0	0.86	0.86
Прийняття інформації через відеоконтент	0	0.84	0.84
Довіра до джерела отримання підробки	0.57	1.9900000000000002	2.56
Перевірка відео на достовірність в інших джерелах	2.55	0.87	3.42
Інформаційно-психологічний вплив на цільову групу осіб (маніпуляція, пропаганда, сіяння паніки тощо)	2.0999999999999996	1.44	3.5399999999999996
Неприйняття змісту підробки	1.0499999999999998	2.39	3.44
Політична криза та (або) зміна вектора державної політики діючої влади	0.88	0	0.88
Ефект праймінгу (priming effect)	0.73	1.4	2.13

Рисунок 3.3 – Ключові показники нечіткої когнітивної карти

Внесок окремого концепту в загальну структуру когнітивної карти можна оцінити шляхом обчислення його загальної центральності (див. стовпець «Centrality» на Рисунку 3.3). Загальна центральність визначається як сума вхідної

та вихідної центральностей для конкретного концепту. Цей показник відображає рівень пов’язаності концепту з іншими концептами моделі, а також сумарну силу відповідних зв’язків.

З аналізу наведених показників видно, що найменший вплив у межах розробленої моделі має концепт «Прийняття інформації через відеоконтент», оскільки його загальна центральність становить 0,84, що є найнижчим значенням серед усіх концептів карти.

Додатковий аналіз показників дозволяє зробити висновок, що основним отримувачем впливу в моделі виступає концепт «Політична криза та (або) зміна вектору державної політики діючої влади», що цілком логічно узгоджується з побудованою структурою моделі.

Component	Indegree	Outdegree	Centrality	Preferred State	Type
Сприймана достовірність підробки	2.95	2.6799999999999997	5.63		ordinary
Недовіра до діючої політичної влади	1.5	1.77	3.27		ordinary
Недовіра до офіційних джерел інформації	2.3	0.39	2.69		ordinary
Якісна підробка Deepfake	0	0.86	0.86		driver
Прийняття інформації через відеоконтент	0	0.84	0.84		driver
Довіра до джерела отримання підробки	0.57	1.9900000000000002	2.56		ordinary
Перевірка відео на достовірність в інших джерелах	2.55	0.87	3.42		ordinary
Інформаційно-психологічний вплив на цільову групу осіб (маніпуляція, пропаганда, сіяння паніки тощо)	2.0999999999999996	1.44	3.5399999999999996		ordinary
Неприйняття змісту підробки	1.0499999999999998	2.39	3.44		ordinary
Політична криза та (або) зміна вектора державної політики діючої влади	0.88	0	0.88		receiver
Ефект праймінгу (priming effect)	0.73	1.4	2.13		ordinary

Рисунок 3.4 – Концепт, що є головним отримувачем (англ. receiver) впливу

Застосування сценарного моделювання дає змогу проаналізувати, як змінюється вплив концептів між собою, і формувати прогностичні оцінки розвитку досліджуваної ситуації.

1. Проаналізуємо, як вплине на стан системи підвищення або зниження значення С7 «Перевірка відео на достовірність в інших джерелах».

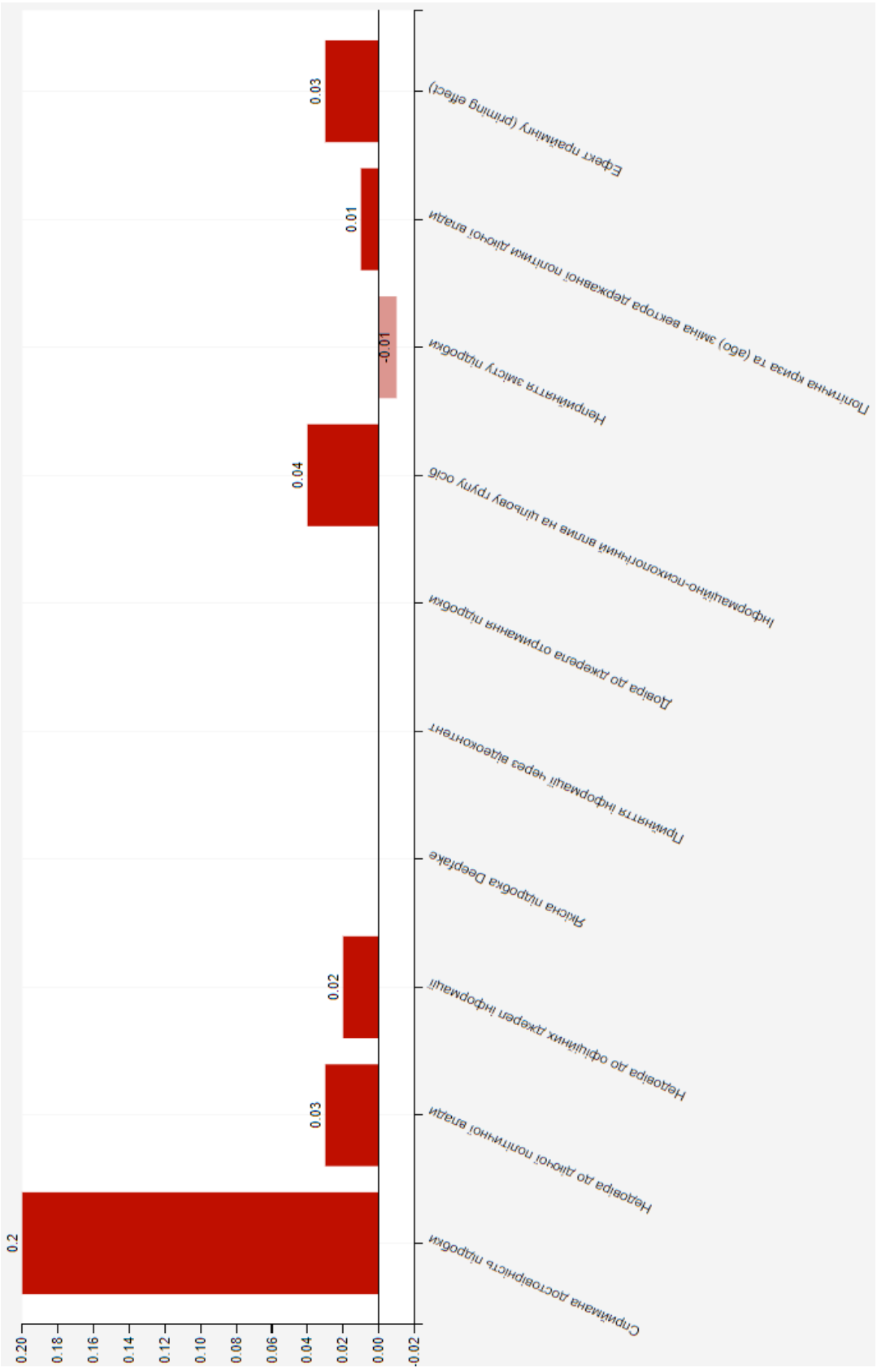

Рисунок 3.5 – Реакція моделі на максимальну негативну зміну значення концепту С7

Рисунок 3.6 – Реакція моделі на максимальну позитивну зміну значення концепту C7

Аналіз отриманих гістограм дозволяє визначити, які саме концепти найбільше піддаються впливу внаслідок зміни значення концепту «Перевірка відео на достовірність в інших джерелах». Зокрема, при максимальному зменшенні значення цього концепту (рис. 3.5) спостерігається суттєвий вплив на концепт «Сприймана достовірність підробки» – його значення зростає на 0,2. Крім того, незначне збільшення показників фіксується у концептів «Недовіра до діючої політичної влади», «Недовіра до офіційних джерел інформації», «Інформаційно-психологічний вплив на цільову групу осіб», «Політична криза та (або) зміна вектору державної політики діючої влади» та «Ефект праймінгу (priming effect)». Водночас дещо зменшується значення концепту «Неприйняття змісту підробки».

У випадку максимального збільшення значення цього концепту також найбільша зміна спостерігається в концепті «Сприймана достовірність підробки» – його значення знижується на 0,13. Окрім цього, фіксується незначне зниження значень концептів «Недовіра до діючої політичної влади», «Недовіра до офіційних джерел інформації», «Інформаційно-психологічний вплив на цільову групу осіб» та «Ефект праймінгу (priming effect)».

2. Розглянемо, яким чином змінюється стан системи за умови одночасного зростання значення концептів «Якісна підробка Deepfake» та «Довіра до джерела отримання підробки», й зниження значення «Перевірка відео на достовірність в інших джерелах», що виступає логічним наслідком підвищення перших двох.

Можемо спостерігати (див. рис. 3.7) значний вплив такої комбінації змін на модель. Найбільший вплив фіксується на концепті «Сприймана достовірність підробки», приріст значення якого становить 0,25. Крім того, незначні зміни спостерігаються у концептів «Недовіра до діючої політичної влади», «Недовіра до офіційних джерел інформації», «Інформаційно-психологічний вплив на цільову групу осіб», «Політична криза та (або) зміна вектору державної політики діючої влади», «Ефект праймінгу (priming effect)» та «Неприйняття змісту підробки». Водночас варто зазначити, що масштаб цих змін є відносно незначним у порівнянні з впливом на основний концепт.

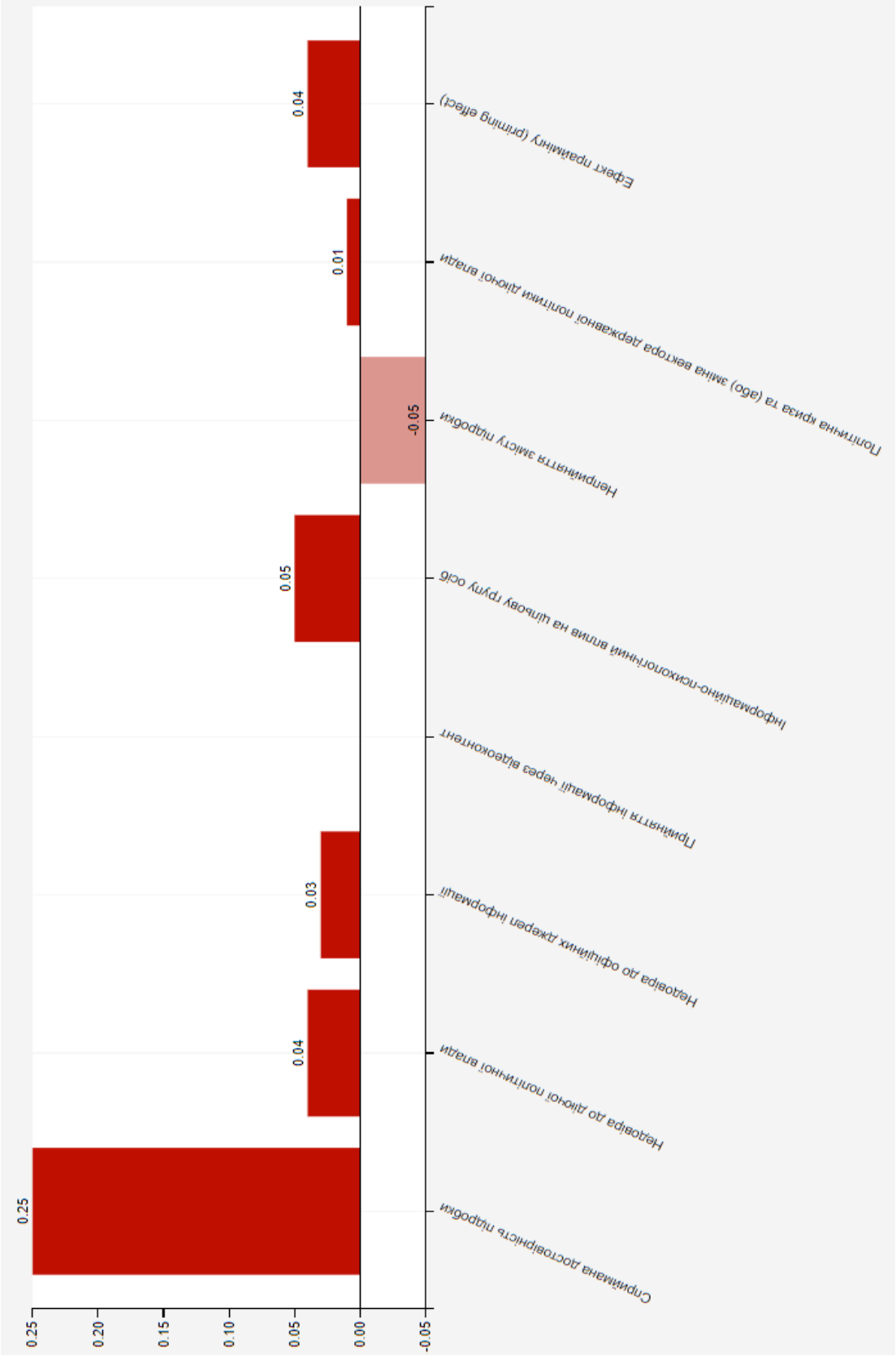

Рисунок 3.7 – Реакція моделі на максимальну позитивну зміну значень концептів С4 та С6, й негативну для С7

3. Проаналізуємо зміни стану системи в разі одночасного підвищення значень концептів «Інформаційно-психологічний вплив на цільову групу осіб», «Ефект праймінгу (priming effect)», й «Недовіра до діючої політичної влади».

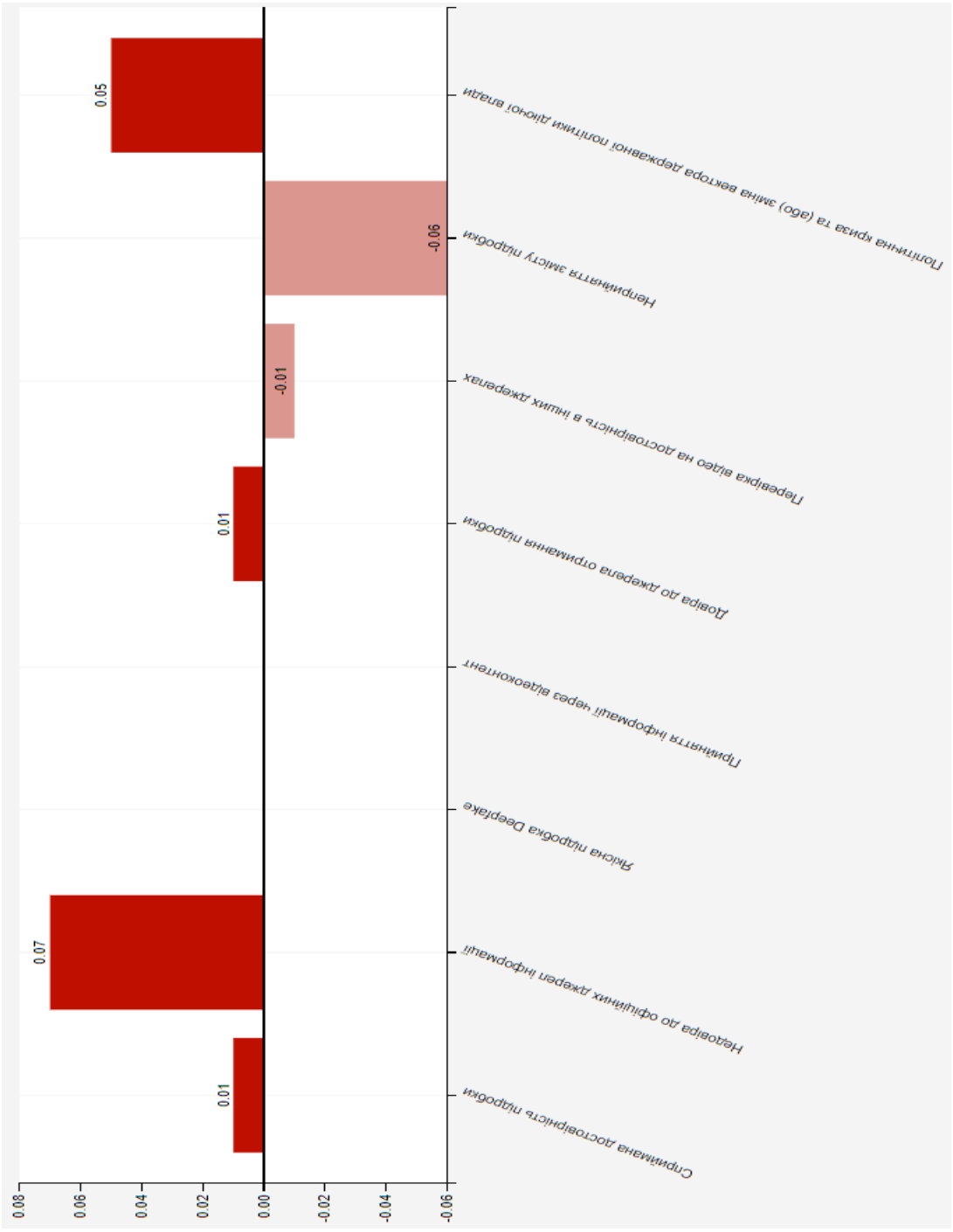

Рисунок 3.8 – Реакція моделі на максимальну позитивну зміну значень концептів С2, С8 та С11

Аналіз отриманої гістограми демонструє вплив зазначених умов на стан моделі. Найбільш істотні зміни спостерігаються серед концептів, що безпосередньо пов’язані з тими, значення яких були змінені. Найбільше зростання показали концепти «Недовіра до офіційних джерел інформації» та «Політична криза та (або) зміна вектору державної політики діючої влади» – їх значення зросли на 0,07 та 0,05 відповідно. Негативна динаміка спостерігається для концепту «Неприйняття змісту підробки», де зафіксовано зменшення на 0,06. Водночас концепти «Сприймана достовірність підробки», «Довіра до джерела отримання підробки» та «Перевірка відео на достовірність в інших джерелах» зазнали лише незначних змін у порівнянні з початковим станом моделі.

Таким чином, результати моделювання нечіткої когнітивної карти дають змогу здійснювати аналіз і прогнозування впливу Deepfake на інформаційну безпеку держави. Сценарне моделювання допомагає досліджувати взаємозв’язки між концептами, включаючи приховані впливи. Отримані висновки доцільно враховувати під час розробки заходів з протидії Deepfake, який може бути застосований у цілях інформаційно-психологічного впливу.

1.2 Модель впливу комплексної протидії Deepfake

Після першого етапу моделювання маємо допрацювати модель – включити в розрахунок комплексну протидію Deepfake, яку будуть використовувати актори кібервійн для зменшення та/або усунення ризиків, що пов’язані з загрозами Deepfake. Отже, на вже існуючу модель маємо додати концепти, що відповідають за протидію, а також встановити дуги – взаємовпливи з вже існуючими концептами, а також ваги дуг в результаті експертного обговорення.

Найголовнішим завданням на цьому етапі є визначення ролі програмно-технічних засобів виявлення та валідації підробок Deepfak, які є ключовими для доведеної та обґрунтованої впевненості у підробці. Саме технічним методам виявлення підробок Deepfake присвячується багато наукових робіт сьогодення. Проте, відповідних статистичних даних про використання широкими народними

масами детекторів Deepfake, а також детекторів контенту, що створений за допомогою генеративних нейронних мереж, більш простих у порівнянні з Deepfake-технологіями, знайдено не було в процесі дослідження тематики.

Більш того, дані, які стосуються шахрайського використання зображень, створених генеративними нейронними мережами [32], свідчать про протилежне – абсолютна більшість користувачів соцмереж покладаються на власні вміння розпізнавання штучного контенту та не використовують навіть застосунки-детектори контенту, що згенерований генеративними нейронними мережами, на кшталт «Midjourney», що виконані в архітектурі глибоких генеративних нейронних мереж, що генерують зображення виходячи з наданого користувачем текстового опису. В інакшому випадку шахрайство з використанням такого контенту не було б так широко розповсюджене в соціальних мережах. Не оминув цей світовий тренд і Україну – у 2024 в соціальній мережі Facebook набули значного поширення публікації зі згенерованими зображеннями (інколи відео), у яких автори дописів просили читачів якомога більше взаємодіяти з публікацією: ставити лайки, коментувати, робити репости тощо [32]. Автори таких дописів використовували чутливі для українців теми, як от війна та ставлення до військових [34], для привернення уваги, а поширення набуло таких розмахів, що ситуацію довелося коментувати навіть Центру Протидії Дезінформації при РНБО України [35]. Фахівці центру висунули два варіанти для чого це робиться, по-перше такі дописи «збирають» на сторінку підписників, які не надто критично ставляться до контенту в мережі та легше за інших сприймають інформаційні маніпуляції, що робить їх вразливими для подальшого впливу, зокрема для російських інформаційно-психологічних операцій, а по-друге з метою шахрайства, по-типу фішингу. Ще одним можливим застосуванням є підвищення таким чином статистичних показників сторінки, що тягне за собою збільшення виплат від Facebook за рекламу. Спостерігачі повідомляли, що незважаючи на часто погану якість штучних фото користувачі часто довіряли подібним публікаціям [36].

Рисунок 3.9 – Приклад публікації зі згенерованим зображенням

А от самостійне виявлення користувачами згенерованих зображень досліджено в багатьох працях – відсоток успішно виявлених фото-фейків варіюється в діапазоні від ~40% до ~70% та залежить від віку користувачів, їх обізнаності в сучасних технологіях генерації зображень, а також якості й фотореалістичності згенерованих матеріалів [37-41].

Отже, маємо часткові докази, що більшість користувачів не використовує детектори Deepfake-підробок, оскільки не використовують навіть детектори контенту, згенерованого генеративними нейромережами, який за замовчуванням

має менш довершене виконання аніж підробки Deepfake. За таких умов програмно-технічні засоби виявлення підробок Deepfake залишаються чудовими інструментами саме для фахівців та не матимуть вирішального значення для зменшення та/або усунення ризиків, що пов'язані з загрозами Deepfake для суспільної думки.

Згрупуємо решту складових комплексної протидії Deepfake у дві групи заходів: превентивні (проактивні) та реактивні.

До превентивних заходів відносяться:

- Просвітницькі кампанії – підвищення медіаграмотності та обізнаності про загрози Deepfake-підробок всього населення або конкретних груп може сприяти підвищенню недовіри до проявів Deepfake та зменшувати його інформаційно-психологічний влив.
- Превентивна комунікація – регулярне інформування населення про можливі загрози Deepfake ще до виникнення криз, що готує громадян до правильного сприйняття інформації. Яскравим прикладом превентивного інформування є дії Головного управління розвідки Міністерства оборони України 3 березня 2022-го року – тоді базуючись на розвідувальних даних розвідники попередили, що Російська Федерація має намір вкинути відео, що буде підробкою Deepfake за участю Президента України Володимира Зеленського, головною метою якого мало б бути сіяння паніки, дезорієнтації, та зневіри серед українських військовослужбовців [42]. Такий матеріал дійсно був опублікований 16 березня 2022 та потрапив до медіа-простору України – у Deepfake-підробці Володимир Зеленський нібито закликав українських військових «скласти зброю та повернутися до своїх сімей». Експерт з «BBC Monitoring» висловлювали думку, що це був один з найгірших Deepfake за технічним виконанням з поміж тих, що він бачив [43] – див. рис 3.10.

У свою чергу до реактивних заходів можна віднести такі:

- Юридична відповідальність за поширення Deepfake – удосконалення законодавства для визначення злочинів, пов'язаних із поширенням deepfake

(наприклад, політичні маніпуляції чи наклеп), із зобов'язаннями для адміністрацій конкретних сторінок, соціальних мереж і медіаплатформ видаляти виявлені deepfake, а також чіткими санкціями за недотримання.

- Спростування – найпростіший метод реакції на появу Deepfake в медіапросторі. Може включати як офіційні спростування, так і поширення спростування в популярних ЗМІ, співпрацю з впливовими блогерами тощо.

Рисунок 3.10 – Порівняння кадрів з підробки Deepfake від 16.03.2022 та реального виступу Володимира Зеленського

Тож, доповнимо множину концептів наступними концептами, які при моделюванні відповідають за комплексну протидію:

1. C12 «Офіційне спростування та в популярних медіа»;
2. C13 «Спростування у джерелі отримання Deepfake» – подібне спростування є дещо ефективнішим, наприклад коли джерело опублікувало підробку ненавмисно, то спростування у цьому ж джерелі швидше переконає користувача у неправдивості змісту підробки [44];
3. C14 «Юридична відповідальність за поширення Deepfake»;

4. C15 «Просвітницькі кампанії та превентивна комунікація».

Як результат, оновлена таблиця концептів нечіткої когнітивної карти виглядатиме наступним чином – див. табл. 3.4.

Таблиця 3.4 – Обрані концепти нечіткої когнітивної карти

Коротке позначення	Концепт
C1	Сприймана достовірність підробки
C2	Недовіра до діючої політичної влади
C3	Недовіра до офіційних джерел інформації
C4	Якісна підробка Deepfake
C5	Прийняття інформації через відеоконтент
C6	Довіра до джерела отримання підробки
C7	Перевірка відео на достовірність в інших джерелах
C8	Інформаційно-психологічний вплив на цільову групу осіб
C9	Неприйняття змісту підробки
C10	Політична криза та (або) зміна вектору державної політики діючої влади
C11	Ефект праймінгу (priming effect)
C12	Офіційне спростування та в популярних медіа
C13	Спростування у джерелі отримання Deepfake
C14	Юридична відповідальність за поширення Deepfake
C15	Просвітницькі кампанії та превентивна комунікація

Тож, тепер можемо додати концепти комплексної протидії на графічну частину нечіткої когнітивної карти. Результат наведено нижче на Рисунку 3.11, що містить множини вершин (концептів) та направлених дуг графа, додані концепти комплексної протидії позначені помаранчевим кольором.

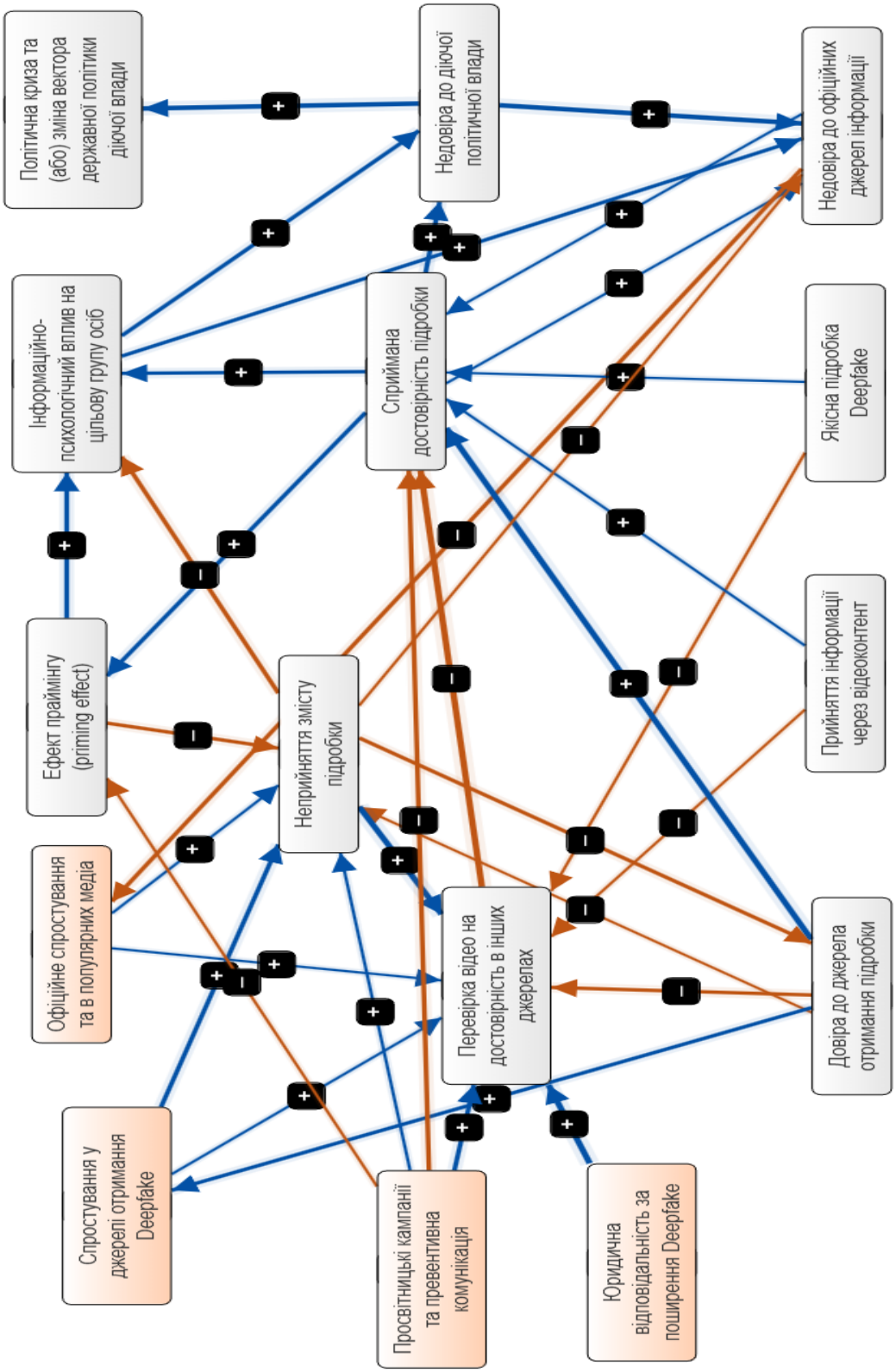

Рисунок 3.11 – Нечітка когнітивна карта для дослідження впливу компексної протидії Deepfake

Тепер, подібно до першого етапу моделювання, експертна комісія узгодила спочатку лінгвістичні значення взаємовпливів концептів, після чого – вже

числові значення вагів взаємовпливів концептів. Для нових концептів наведемо значення їх взаємовпливів з іншими концептами моделі на Рисунках 3.12-3.13, а також та повну матрицю взаємовпливів концептів на Рисунку 3.14.

	Сприймана достовірність підробки	Недовіра до діючої політичної влади	Недовіра до офіційних джерел інформації	Якісна підробка Deepfake	Прийняття інформації через відеоконтент	Довіра до джерела отримання підробки	Перевірка відео на достовірність в інших джерелах	Інформаційно-психологічний вплив на цільову групу осіб	Неприйняття змісту підробки	Політична криза та (або) зміна вектора державної політики діючої влади	Ефект праймінгу (priming effect)
Офіційне спростування та в популярних медіа							0.34		0.43		
Спростування у джерелі отримання Deepfake							0.46		0.74		
Юридична відповідальність за поширення Deepfake							0.87				
Просвітницькі кампанії та превентивна комунікація	-0.72						0.66		0.44		-0.47

Рисунок 3.12 – Впливи нових концептів (вертикальна вісь) на інші концепти (горизонтальна вісь)

	Офіційне спростування та в популярних медіа	Спростування у джерелі отримання Deepfake	Юридична відповідальність за поширення Deepfake	Просвітницькі кампанії та превентивна комунікація
Сприймана достовірність підробки				
Недовіра до діючої політичної влади				
Недовіра до офіційних джерел інформації	-0.78			
Якісна підробка Deepfake				
Прийняття інформації через відеоконтент				
Довіра до джерела отримання підробки		0.81		

Рисунок 3.13 – Впливи інших концептів (вертикальна вісь) на нові концепти (горизонтальна вісь)

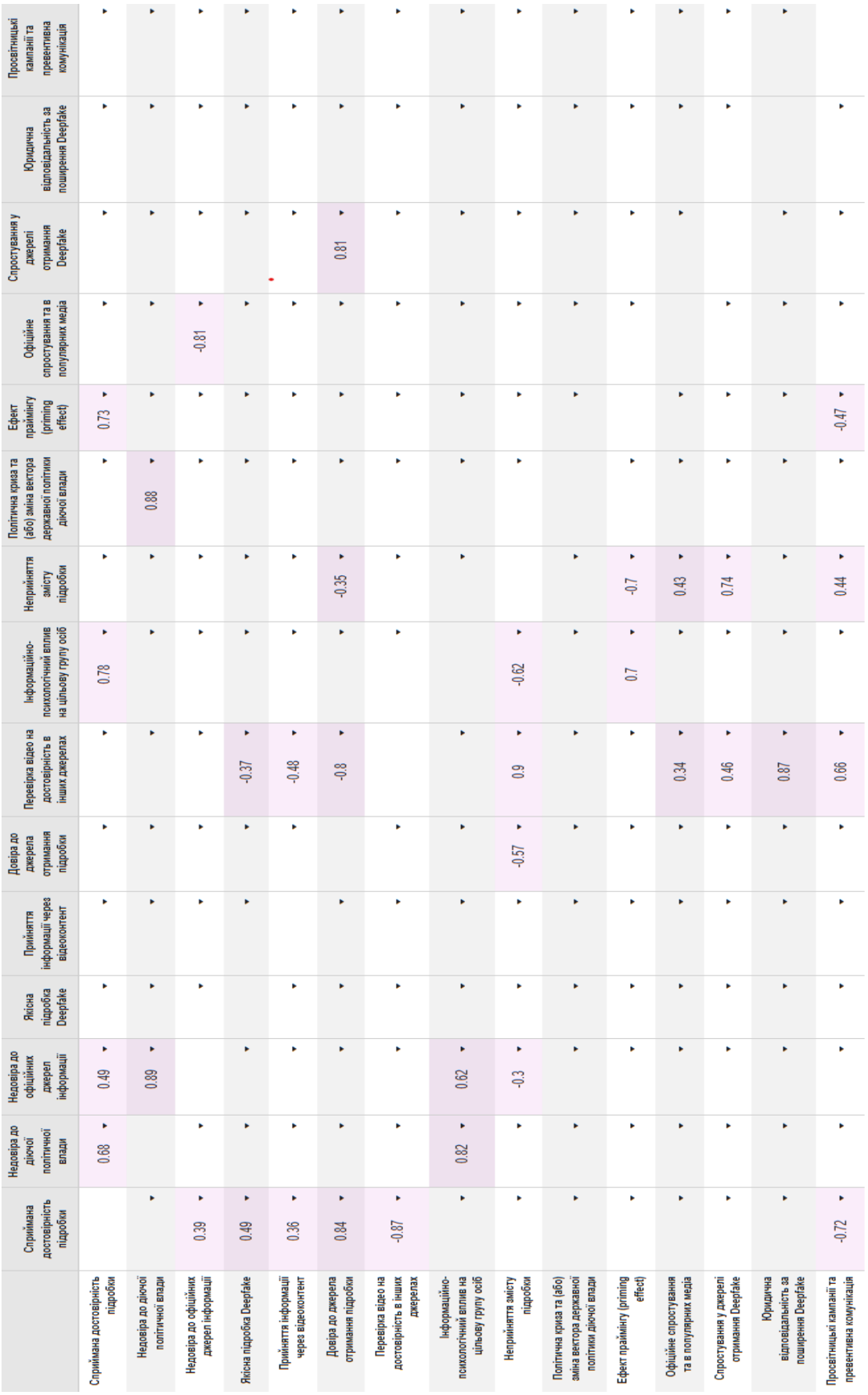

Рисунок 3.14 – Матриця взаємовпливів концептів НКК

Визначимо зв’язність розробленої нечіткої когнітивної карти за допомогою коефіцієнта кластеризації, який дасть змогу оцінити щільність зв’язків представленого графа.

Для цього використаємо формулу (3.5). Маємо дані, що кількість концептів N = 15, кількість всіх зв’язків графа M = 34.

Підставимо ці значення у формулу (3.5) – отримуємо, що коефіцієнт кластеризації D = $\frac{34}{210}$ = 0,1619.

Таке числове значення менше за коефіцієнт кластеризації попереднього етапу моделювання, що вказує на меншу зв’язність моделі на другому етапі моделювання, проте таке значення все ще вказує на достатню зв’язність нечіткої когнітивної карти.

Це пояснюється тим, що внаслідок еволюції моделі зменшилося число щільних груп концептів. Оскільки самих концептів стало більше на ~36,4%, а дуг взаємовпливів на ~47,8% – виходячи з представлення формули 3.5 бачимо, що приріст кількості взаємовпливів недостатній, щоб компенсувати збільшення кількості концептів.

Варто зазначити, що у випадку, коли щільність карти висока – це свідчитиме про наявність великої кількості причинних зв'язків між змінними. Що більше в карті взаємозв'язків, то більше є можливостей для зміни ситуації. Проте, якщо карта має меншу щільність – тобто між концептами існує нормована кількість зв’язків, то це значить, що карта є не відчутно складною для розуміння та потенційних змін ситуації.

На Рисунку 3.15 наведено розраховані основні кількісні показники розробленої нечіткої когнітивної карти впливу комплексної протидії Deepfake.

Загальна центральність концепту (на Рисунку 3.15 стовпець «Centrality»), що обчислюється сумою вхідної і вихідної центральності для нього, описує вклад концепту у когнітивну карту, показує, як він пов'язаний з іншими концептами та яка сукупна сила цих зв'язків. З наведених вище показників видно, що найменший вплив на роботу розробленої моделі має концепт «Прийняття

інформації через відеоконтент» так само, як і на попередньому етапі моделювання, його загальна центральність складає 0,84 , тобто залишається найменшим показником центральності для всіх концептів моделі.

Total Components	15
Total Connections	34
Density	0.1619047619
Connections per Component	2.2666666667
Number of Driver Components	4
Number of Receiver Components	1
Number of Ordinary Components	10
Complexity Score	0.25

Component	Indegree	Outdegree	Centrality	Preferred State	Type
Сприймана достовірність підробки	3.67	2.6799999999999997	6.35		ordinary
Недовіра до діючої політичної влади	1.5	1.77	3.27		ordinary
Недовіра до офіційних джерел інформації	2.3	1.17	3.4699999999999998		ordinary
Якісна підробка Deepfake	0	0.86	0.86		driver
Прийняття інформації через відеоконтент	0	0.84	0.84		driver
Довіра до джерела отримання підробки	0.57	2.8000000000000003	3.37		ordinary
Перевірка відео на достовірність в інших джерелах	4.88	0.87	5.75		ordinary
Інформаційно-психологічний вплив на цільову групу осіб	2.0999999999999996	1.44	3.5399999999999996		ordinary
Неприйняття змісту підробки	2.6599999999999997	2.39	5.05		ordinary
Політична криза та (або) зміна вектора державної політики діючої влади	0.88	0	0.88		receiver
Ефект праймінгу (priming effect)	1.2	1.4	2.5999999999999996		ordinary
Офіційне спростування та в популярних медіа	0.78	0.77	1.55		ordinary
Спростування у джерелі отримання Deepfake	0.81	1.2	2.01		ordinary
Юридична відповідальність за поширення Deepfake	0	0.87	0.87		driver
Просвітницькі кампанії та превентивна комунікація	0	2.29	2.29		driver

Рисунок 3.15 – Основні показники розробленої нечіткої когнітивної карти

Визначено, що головним отримувачем впливу залишається концепт «Політична криза та (або) зміна вектору державної політики діючої влади», що логічно витікає з архітектури розробленої моделі. Змін з попереднього етану моделювання не помічено.

Component	Indegree	Outdegree	Centrality	Preferred State	Type
Сприймана достовірність підробки	3.67	2.6799999999999997	6.35		ordinary
Недовіра до діючої політичної влади	1.5	1.77	3.27		ordinary
Недовіра до офіційних джерел інформації	2.3	1.17	3.4699999999999998		ordinary
Якісна підробка Deepfake	0	0.86	0.86		driver
Прийняття інформації через відеоконтент	0	0.84	0.84		driver
Довіра до джерела отримання підробки	0.57	2.8000000000000003	3.37		ordinary
Перевірка відео на достовірність в інших джерелах	4.88	0.87	5.75		ordinary
Інформаційно-психологічний вплив на цільову групу осіб	2.0999999999999996	1.44	3.5399999999999996		ordinary
Неприйняття змісту підробки	2.6599999999999997	2.39	5.05		ordinary
Політична криза та (або) зміна вектора державної політики діючої влади	0.88	0	0.88		receiver
Ефект праймінгу (priming effect)	1.2	1.4	2.5999999999999996		ordinary
Офіційне спростування та в	0.78	0.77	1.55		ordinary

Рисунок 3.16 – Концепт, що є головним отримувачем (англ. receiver) впливу від решти концептів нечіткої когнітивної карти

Складність нечіткої карти (англ. complexity score) – це співвідношення кількості головних концептів-отримувачів впливу (англ. receiver) до концептів-драйверів впливу (англ. driver). Складні карти мають велике значення цього коефіцієнта, оскільки в них передбачається більше корисних результатів, що виробляються і використовуються в системі. У нашому випадку маємо один концепт-отримувач, як вказано вище, та чотири концепти-драйвери впливу, а отже складність обраховується як 1/4 = 0,25 (див. рис. 3.17). Такий показник вказує на недостатню складність моделі.

Натомість, у карті переважають «звичайні» (англ. ordinary) концепти, тобто ті, що поєднують у собі функції драйверів і отримувачів впливу - їх налічується 10 від загальної кількості 15-ти концептів (див.рис. 3.18).

Total Components	15
Total Connections	34
Density	0.1619047619
Connections per Component	2.2666666667
Number of Driver Components	4
Number of Receiver Components	1
Number of Ordinary Components	10
Complexity Score	0.25

Component	Indegree	Outdegree	Centrality	Preferred State	Type
Якісна підробка Deepfake	0	0.86	0.86		driver
Прийняття інформації через відеоконтент	0	0.84	0.84		driver
Юридична відповідальність за поширення Deepfake	0	0.87	0.87		driver
Просвітницькі кампанії та превентивна комунікація	0	2.29	2.29		driver
Політична криза та (або) зміна вектора державної політики діючої влади	0.88	0	0.88		receiver

Рисунок 3.17 – Драйвери впливу (синім кольором), отримувач впливу (фіолетовим) та показник складності карти (червоним)

Total Components	15
Total Connections	34
Density	0.1619047619
Connections per Component	2.2666666667
Number of Driver Components	4
Number of Receiver Components	1
Number of Ordinary Components	10
Complexity Score	0.25

Component	Indegree	Outdegree	Centrality	Preferred State	Type
Сприймана достовірність підробки	3.67	2.6799999999999997	6.35		ordinary
Недовіра до діючої політичної влади	1.5	1.77	3.27		ordinary
Недовіра до офіційних джерел інформації	2.3	1.17	3.4699999999999998		ordinary
Довіра до джерела отримання підробки	0.57	2.8000000000000003	3.37		ordinary
Перевірка відео на достовірність в інших джерелах	4.88	0.87	5.75		ordinary
Інформаційно-психологічний вплив на цільову групу осіб	2.0999999999999996	1.44	3.5399999999999996		ordinary
Неприйняття змісту підробки	2.6599999999999997	2.39	5.05		ordinary
Ефект праймінгу (priming effect)	1.2	1.4	2.5999999999999996		ordinary
Офіційне спростування та в популярних медіа	0.78	0.77	1.55		ordinary
Спростування у джерелі отримання Deepfake	0.81	1.2	2.01		ordinary

Рисунок 3.18 – «Звичайні» (англ. ordinary) концепти карти

Тепер проведемо сценарне моделювання тих самих ситуацій, які досліджувалися на першому етапі моделі. А саме:

1. Розглянемо, як зміниться стан системи при збільшенні та зменшенні значення концепту «Перевірка відео на достовірність в інших джерелах»;

2. Розглянемо, як зміниться стан системи при збільшенні значень концептів «Якісна підробка Deepfake» та «Довіра до джерела отримання підробки» та зменшенні значення концепту «Перевірка відео на достовірність в інших джерелах»;

3. Розглянемо, як зміниться стан системи при збільшенні значень концептів «Ефект праймінгу (priming effect)», «Інформаційно-психологічний вплив на цільову групу осіб» та «Недовіра до діючої політичної влади».

Почнемо з варіанту зміни значення концепту «Перевірка відео на достовірність в інших джерелах» (див. рис. 3.19, 3.20). Аналіз отриманих гістограм дозволяє визначити, які саме концепти зазнають найбільшого впливу в результаті зміни цього концепту. Зокрема, при максимальних негативних змінах (рис. 3.19) спостерігається суттєве зростання значення концепту «Сприймана достовірність підробки» – воно підвищується на 0,31, що істотно перевищує відповідний показник на попередньому етапі моделювання.

Як і на минулому етапі моделі незначно збільшяться значення концептів:

- «Недовіра до діючої політичної влади»;
- «Недовіра до офіційних джерел інформації»;
- «Інформаційно-психологічний вплив на цільову групу осіб»;
- «Політична криза та (або) зміна вектору державної політики діючої влади»;
- «Ефект праймінгу (priming effect)».

Також незначно зменшаться значення концептів «Неприйняття змісту підробки» та «Офіційні спростування та у популярних ЗМІ».

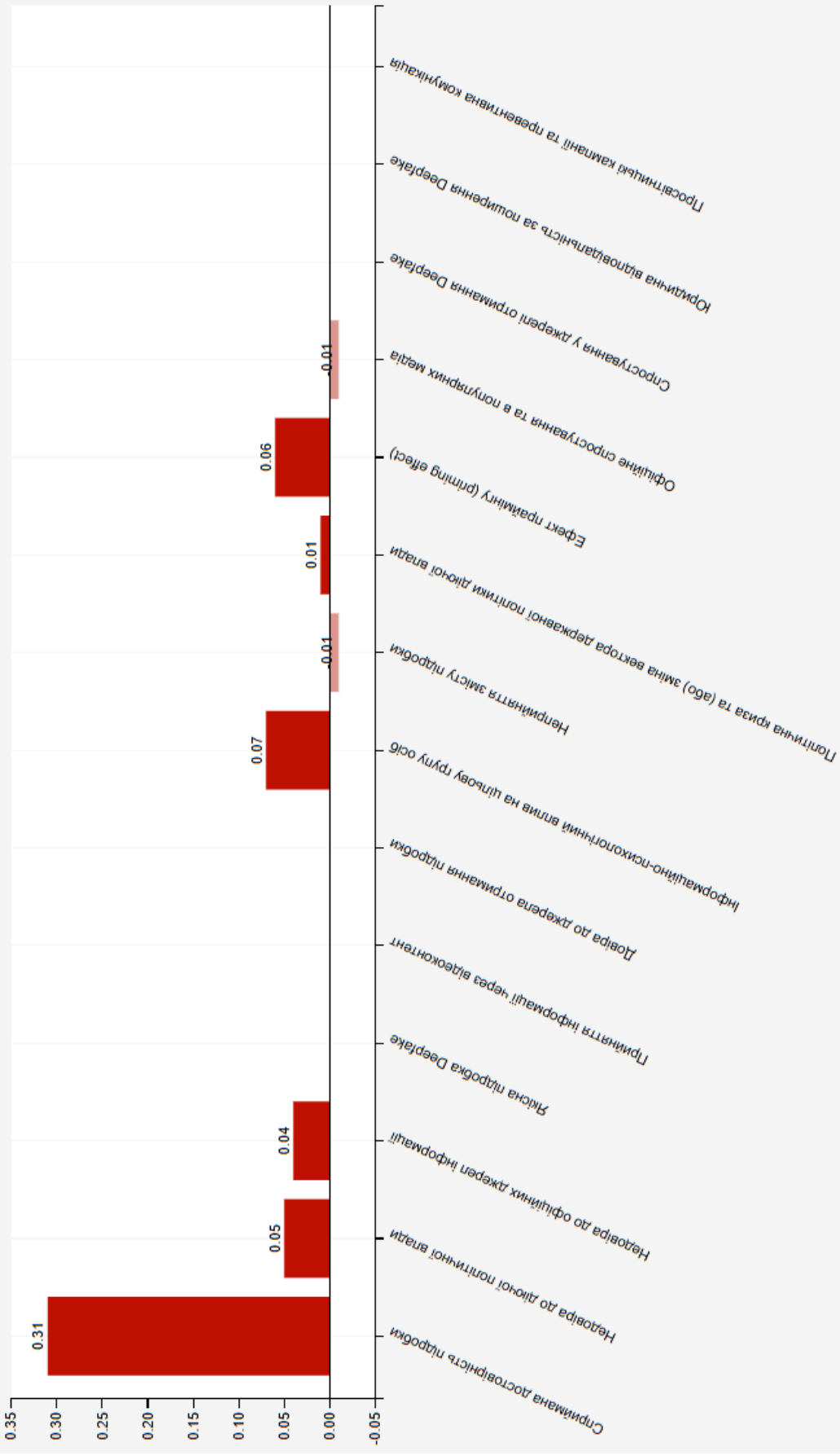

Рисунок 3.19 – Реакція моделі на максимальну негативну зміну значення концепту С7

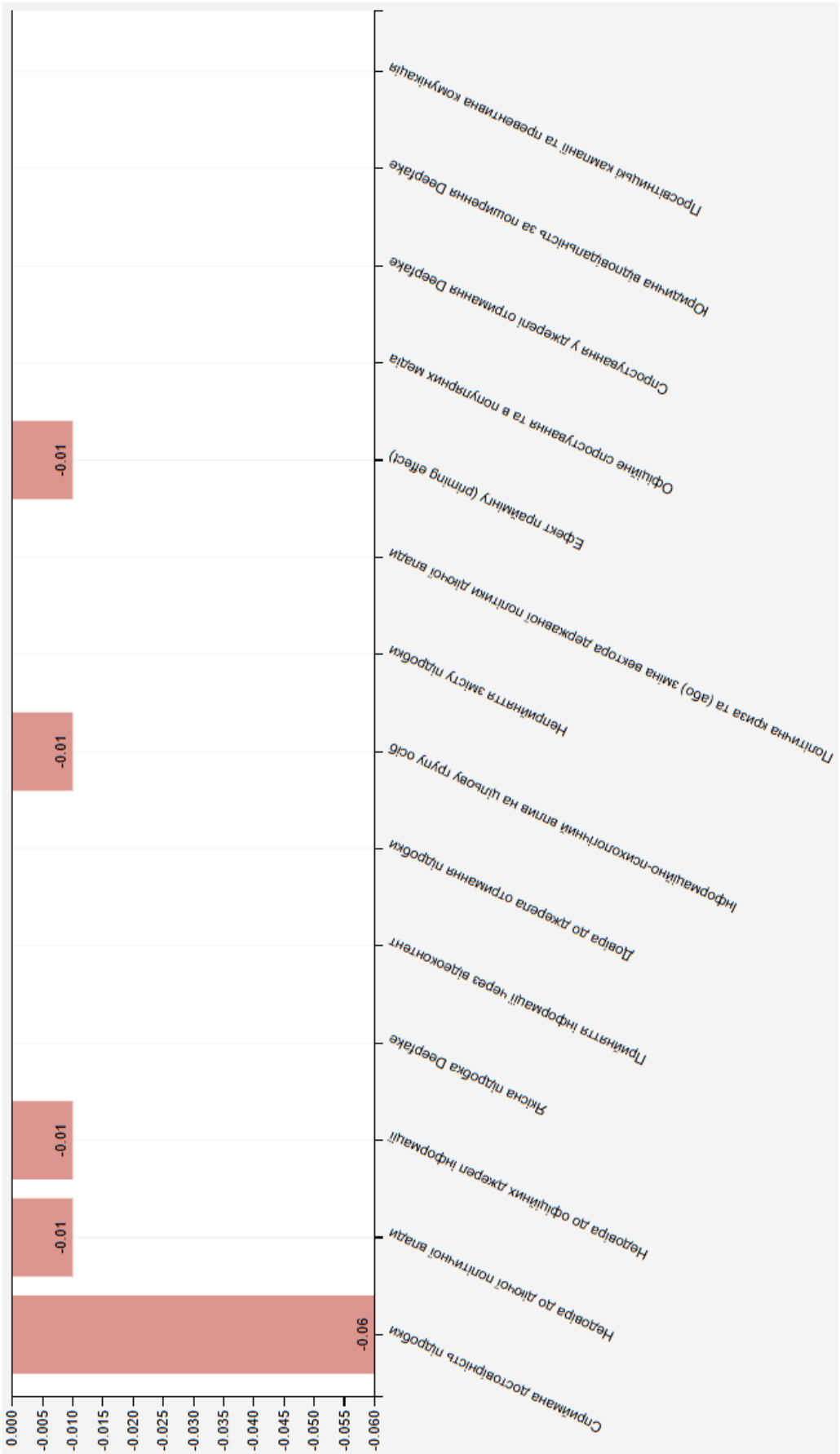

Рисунок 3.20 – Реакція моделі на максимальну позитивну зміну значення концепту C7

У свою чергу, максимальні позитивні зміни , як і в минулому випадку, значно вплинуть на концепт «Сприймана достовірність підробки» – його значення зменшиться на 0,6 – що теж значно відрізняється у меншу сторону порівняно з попереднім етапом. Також незначно зменшиться значення концептів «Недовіра до діючої політичної влади», «Недовіра до офіційних джерел інформації», «Інформаційно-психологічний вплив на цільову групу осіб (маніпуляція, пропаганда, сіяння паніки тощо)» та «Ефект праймінгу (priming effect)».

Такі значущі зміни щодо зміни для концепту «Сприйняття достовірності підробки» можна пояснити тим, що система стала менш зв’язною, а отже менш концепти менше контролюють один одного.

Перевіримо стан системи для сценарію №2 (див. рис. 3.21), що симулюватиме несприятливі умови для протидії Deepfake.

Найбільший вплив припав на концепт «Сприймана достовірність підробки», його значення збільшилося аж на 0,39 – бачимо продовження тенденції з попередніх сценаріїв моделювання, інша структура системи впливає на збільшення змін для цього концепту, який є ключовим транзитним пунктом впливу (він найбільше отримує вплив та найбільше передає вплив).

Також, як і минулого разу, наслідки незначно вплинули на такі концепти : «Недовіра до діючої політичної влади», «Недовіра до офіційних джерел інформації», «Інформаційно-психологічний вплив на цільову групу осіб», «Політична криза та (або) зміна вектору державної політики діючої влади» «Ефект праймінгу (priming effect)», «Неприйняття змісту підробки». Проте їх зміна не значна порівняно з концептом на який припав найбільший вплив. Ще примітна зміна – значення концепту «Спростування у джерелі отримання Deepfake» виросло на 0,11, що є прямим наслідком радикального збільшення довіри до джерела отримання підробки – а отже і спростування у цьому джерелі буде переконливішим.

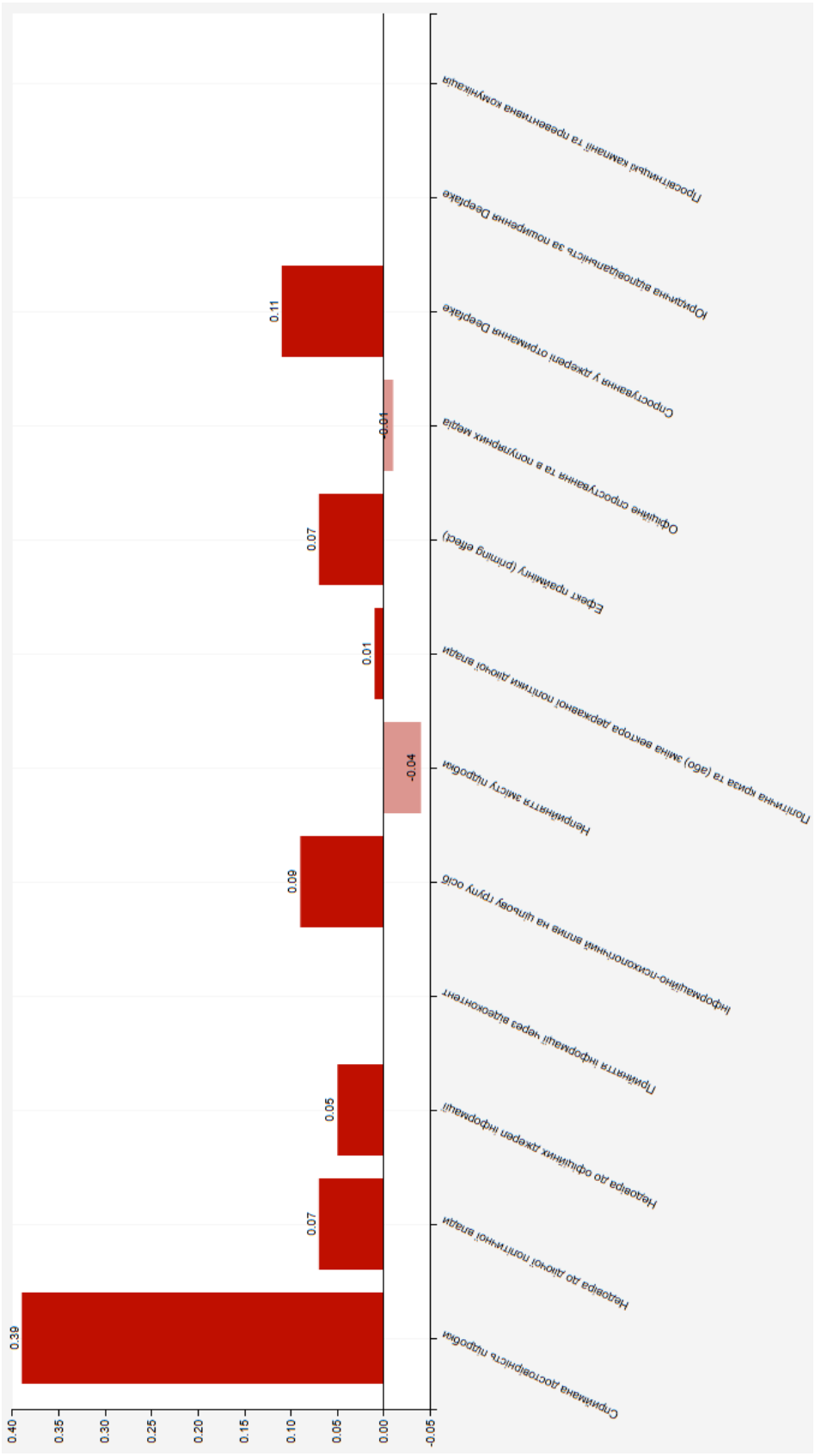

Рисунок 3.21 – Реакція моделі на максимальну позитивну зміну значень концептів С4 та С6, й негативну для С7

Перевіримо стан системи для сценарію №3 (див. рис. 3.22 нижче), що симулюватиме несприятливі умови для протидії Deepfake, проте інші за характером, аніж у попередньому сценарії.

З аналізу отриманої гістограми можна зробити висновок, що змінені умови моделювання найбільше вплинули на ті концепти, які безпосередньо пов’язані з концептами зі зміненими значеннями. Зокрема, найвище зростання продемонстрували концепти «Недовіра до офіційних джерел інформації» та «Політична криза та (або) зміна вектору державної політики діючої влади» – їх значення зросли на 0,09 та 0,06.

Водночас було зафіксовано зменшення значення концепту «Неприйняття змісту підробки» на 0,08, що свідчить про зворотний ефект в межах моделі. Інші концепти – «Сприймана достовірність підробки», «Довіра до джерела отримання підробки», «Перевірка відео на достовірність в інших джерелах» та «Офіційне спростування та в популярних медіа» – продемонстрували мінімальні коливання значень, що свідчить про їхню відносну стійкість до змін у заданих умовах.

Примітно, що значущі зміни відносно першого етапу відсутні – це пояснюється тим, що концепти, яким ми змінювали значення, а також концепти на які це вплинуло досить віддалені по структурі моделі від концептів протидії, які були додані.

Проведемо ще один етап сценарного моделювання, на якому розглянемо, як зміниться стан системи при змінах значення концептів, що відповідають за комплексну протидію технології Deepfake. Спершу збільшимо значення кожного концепту протидії та подивимося на стани системи. А саме почергово збільшимо значення таких концептів (див. рис. 3.23-3.26):

1. С12 «Офіційне спростування та в популярних медіа»;
2. С13 «Спростування у джерелі отримання Deepfake»;
3. С14 «Юридична відповідальність за поширення Deepfake»;
4. С15 «Просвітницькі кампанії та превентивна комунікація».

Після проведення сценарного моделювання порівняємо результати змін, яких зазнала система.

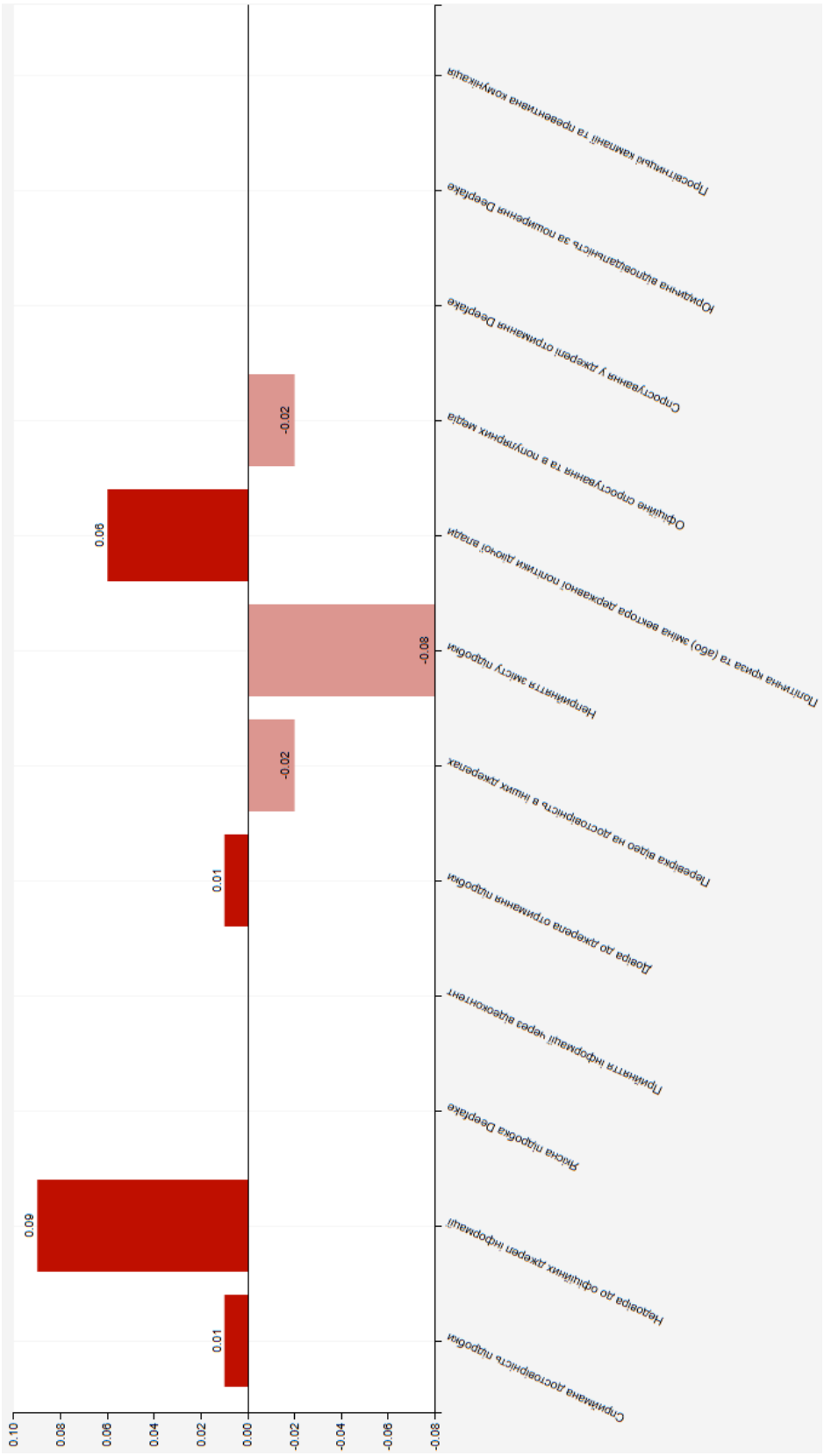

Рисунок 3.22 – Реакція моделі на максимальну позитивну зміну значень концептів C2, C8 та C11

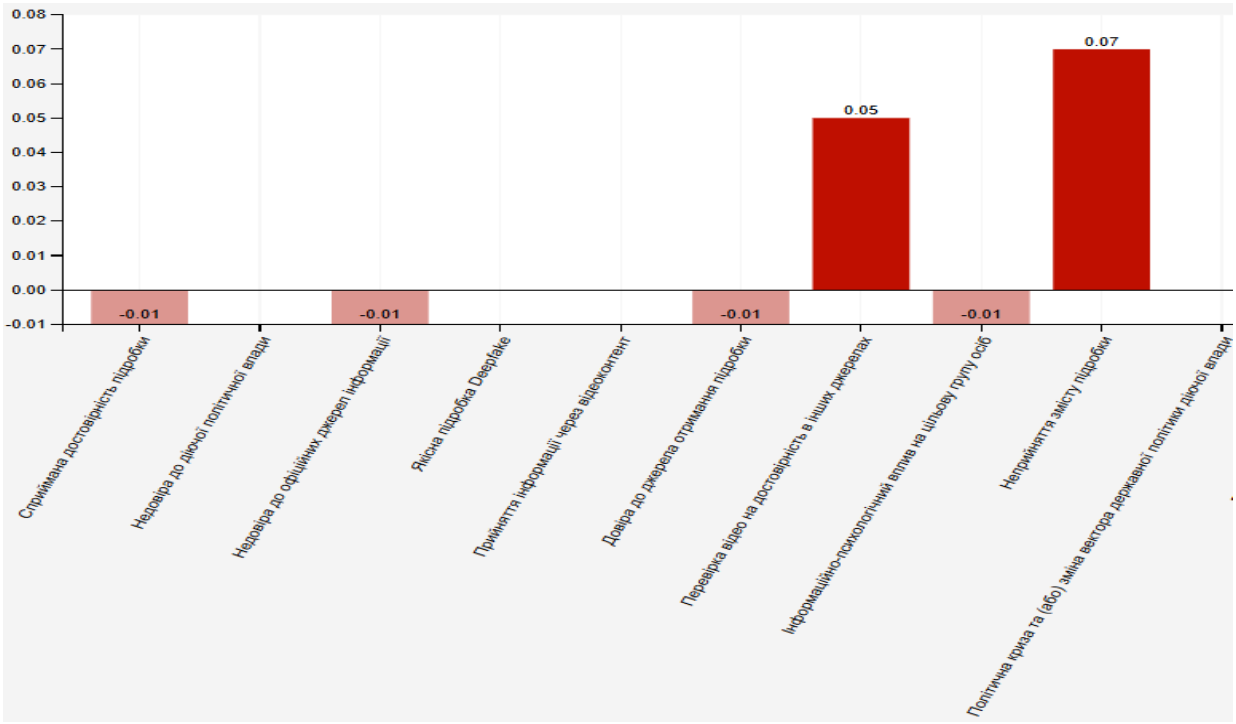

Рисунок 3.23 – Реакція моделі на максимальну позитивну зміну значення концепту С12 «Офіційне спростування та в популярних медіа»

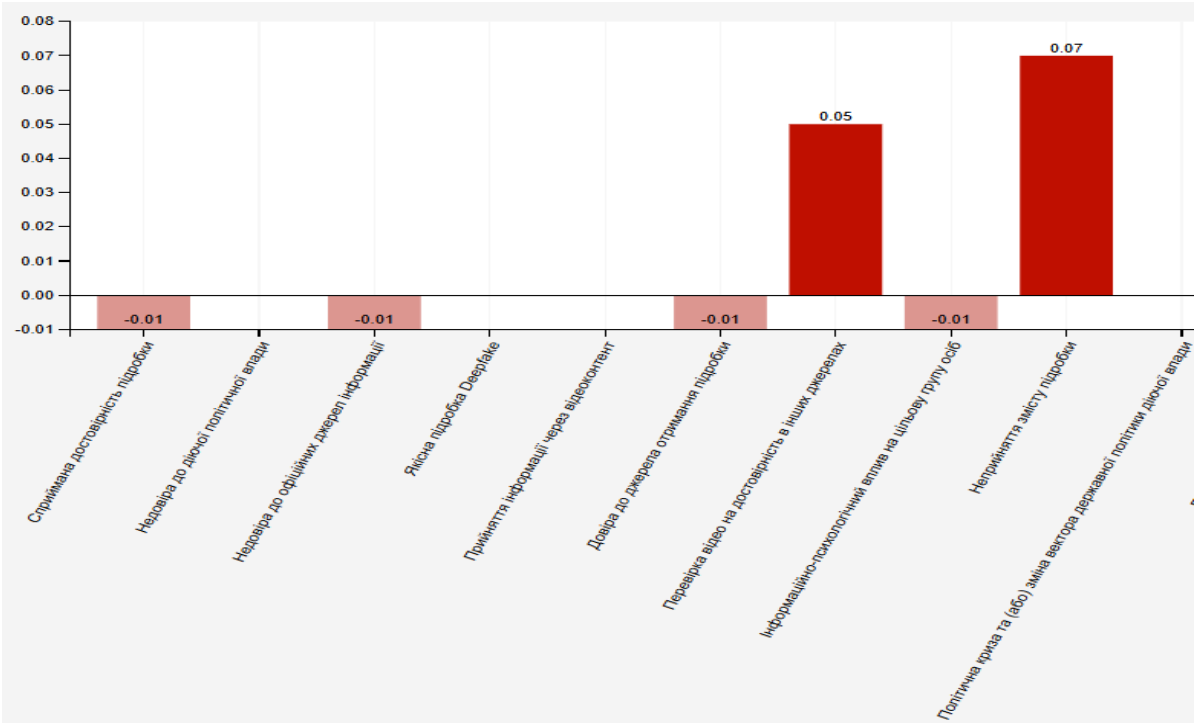

Рисунок 3.24 – Реакція моделі на максимальну позитивну зміну значення концепту С13 «Спростування у джерелі отримання Deepfake»

На отриманих гістограмах, що позначають зміни у значеннях концептів бачимо, що збільшення значень обидвох типів спростування однаково впливають на досліджувану систему. Це пояснюється тим, що обидва спростування мають схожі взаємовпливи з іншими концептами системи, зокрема це видно і на гістограмах – обидва спростування збільшують значення концептів «Неприйняття змісту підробки» та «Перевірка відео на достовірність в інших джерелах». Інші зміни, які вони спричиняють через опосередковані зв'язки, не значні.

Перейдемо до зміни значення концептів, що відповідають за превентивну протидію.

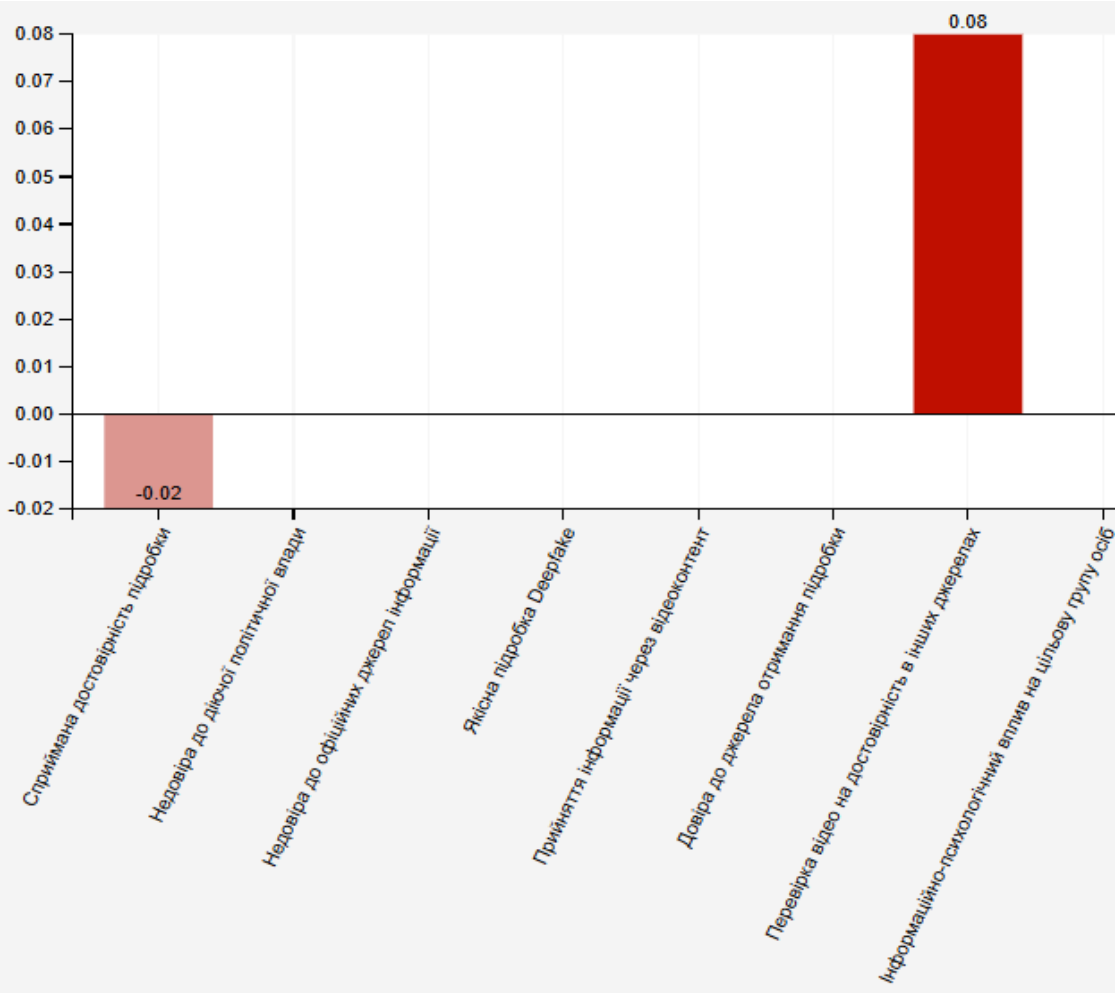

Рисунок 3.25 – Реакція моделі на максимальну позитивну зміну значення концепту С14 «Юридична відповідальність за поширення Deepfake»

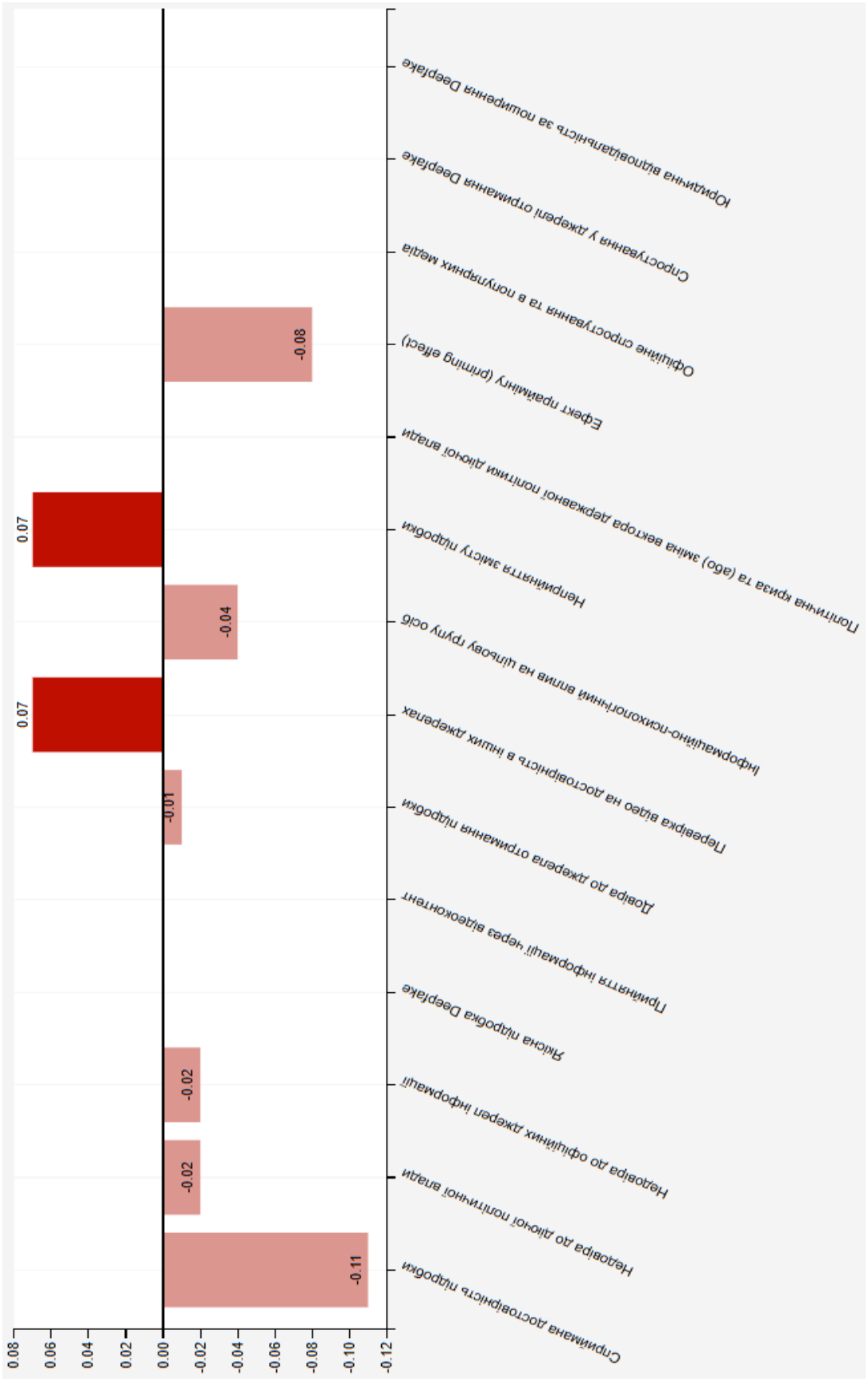

Рисунок 3.26 – Реакція моделі на максимальну позитивну зміну значення концепту С15 «Просвітницькі кампанії та превентивна комунікація»

Бачимо вагомі зміни у стані системи для збільшення значень обидвох типів превентивних дій. Таким чином, максимальне позитивне збільшення значення концепту «Юридична відповідальність за поширення Deepfake» збільшило на 0,08 значення концепту «Перевірка відео на достовірність в інших джерелах», а також опосередковано вплинуло на значення концепту «Сприймана достовірність підробки» – фіксуємо зниження значення на -0,02.

У той же час, максимальна позитивна зміна значення концепту «Просвітницькі кампанії та превентивна комунікація» значно вплинули на 4 концепти: «Сприймана достовірність підробки» – зниження на -0,11, «Ефект праймінгу (priming effect)» – зниження на -0,08, «Інформаційно-психологічний вплив на цільову групу осіб» – зниження на -0,04, а значення концептів «Перевірка відео на достовірність в інших джерелах» та «Неприйняття змісту підробки» збільшилися на 0,07 кожен.

Бачимо, що концепти, які відповідають за превентивну протидію, в цілому краще себе показують у якості засобів протидії – їх вплив відчутніший для системи при зміні значень відповідних концептів.

Далі маємо перевірити зміни системи при максимальних позитивних змінах значень усіх чотирьох концептів комплексної протидії (див. рис. 3.27).

Як і передбачалося при створенні моделі, комплекс заходів протидії значно впливає на систему – при максимальних позитивних змінах значень концептів протидії спостерігаємо збільшення значень концептів, які важливі для протидії Deepfake, а також зменшення значень концептів, які позначають результати та драйвери негативного впливу Deepfake. Зокрема, значення концептів «Перевірка відео на достовірність в інших джерелах» та «Неприйняття змісту підробки» збільшилися на 0,19 кожен. А також помічено такі значні зменшення значень: для концепту «Сприймана достовірність підробки» на -0,14, для «Інформаційно-психологічний вплив на цільову групу осіб» на -0,07, для «Ефект праймінгу (priming effect)» на -0,08, для «Недовіра до офіційних джерел інформації» на -0,4.

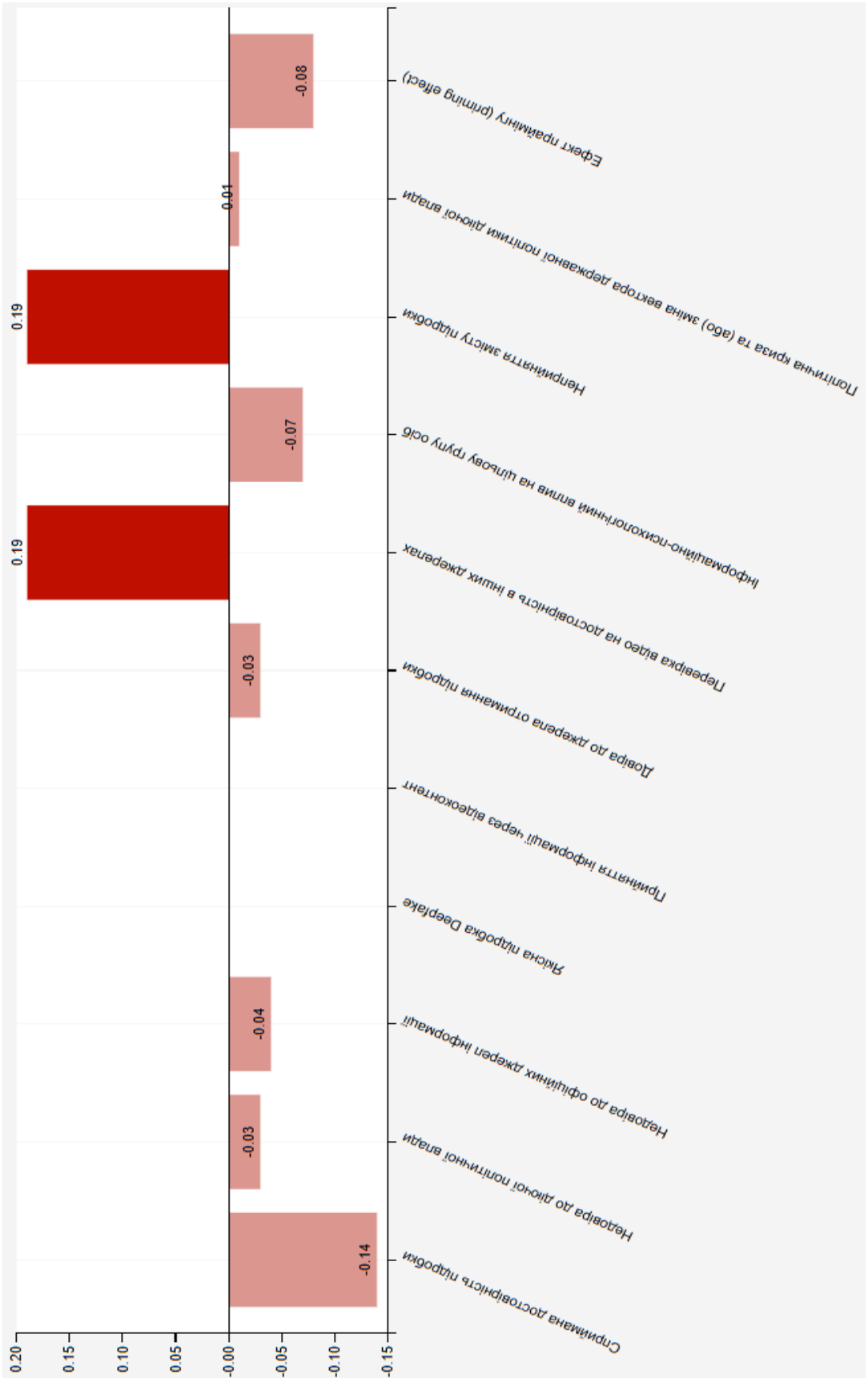

Рисунок 3.27 – Реакція моделі на максимальну позитивну зміну значень концептів C12, C13, C14 та C15

До вищевказаного сценарію додамо ще одну зміну – максимальну позитивну зміну для концепту «Якісна підробка Deepfake», пам’ятаючи, що цей концепт був одним з головних драйверів впливу (див. рис. 3.28).

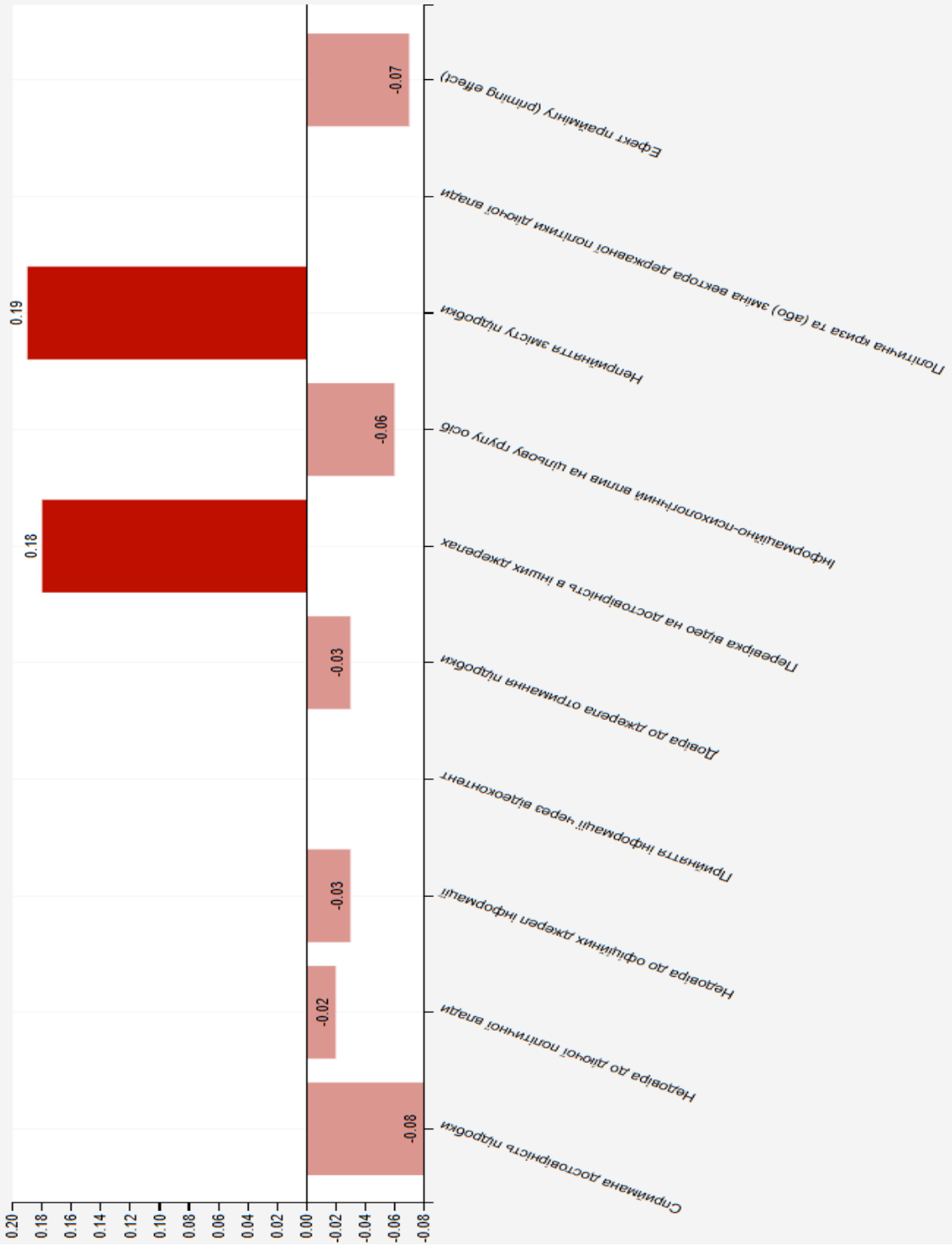

Рисунок 3.28 – Реакція моделі на максимальну позитивну зміну значень концептів C4, C12, C13, C14 та C15

Бачимо результат, збільшене значення концепту-драйверу впливу «Якісна підробка Deepfake» не вплинуло значно на стан системи відносно минулого сценарію, окрім концепту «Сприймана достовірність підробки» – його значення зменшилося лише на -0,08, порівняно з -0,13 при минулому сценарії. Проте, навіть таке зменшення концепту все одно позитивне з точки зору протидії.

Тобто, збільшення значень концептів протидії у даній системі компенсує поодинокі сильні позитині значення «концептів нападу» та все одно зменшує вплив Deepfake.

Висновки до розділу 3

Третій розділ присвячений двом етапам моделювання за допомогою нечіткої логіки, а саме нечітких когнітивних карт. На першому етапі був змодельований вплив Deepfake на інформаційну безпеку держави, ця модель дозволила виявляти приховані зв’язки при впливі Deepfake на інформаційну безпеку держави. Також модель першого етапу дає змогу оцінити значення Deepfake у сучасних інформаційних війнах (у т.ч. в операціях впливу) за допомогою сценарного моделювання потенційного впливу Deepfake на інформаційну безпеку держави. На другому етапі моделювання до вже існуючої моделі додано блок, що відповідає за комплексну протидію Deepfake. Відповідно, можливості виявлення прихованих зав’язків та сценарного моделювання прогнозів впливів значно розширилися. Модель дає змогу досліджувати різні ситуації та розвитки подій як позитивні, так і негативні та виявляти неочевидні на перший погляд наслідки впливів.

ВИСНОВКИ

Тема дослідження, що було виконане, є надзвичайно актуальною наразі, і її важливість лише зростатиме з часом, адже технології нейронних мереж для генерації зображень, відео та аудіозаписів постійно вдосконалюються. Зростання кількості та якості Deepfake-контенту в Інтернеті веде до появи нових ризиків та загроз, що, в свою чергу, стимулює активні дослідження в сфері методів виявлення та протидії Deepfake. У рамках цієї роботи було досліджено сучасний стан проблематики Deepfake та кібервійн, основні методи виробництва підробок Deepfake, приналежність до кіберзброї та місце в методах ведення сучасних кібервійн. За результатами дослідження була розроблена власна класифікація груп методів ведення кібервійн, а також моделі: впливу Deepfake на інформаційну безпеку держави та впливу комплексної протидії Deepfake – після чого було проведене сценарне моделювання з аналізом отриманих результатів.

Практична цінність отриманих результатів полягає в тому, що створена в межах цього дослідження модель впливу на інформаційну безпеку держави та її розширення – модель впливу комплексної протидії – можуть бути використані для подальшого аналізу негативного ефекту Deepfake та вдосконалення підходів до протидії їх поширенню й впливу. Зокрема, вони можуть слугувати основою для проведення сценарного моделювання сприйняття Deepfake-підробок та комплексних заходів протидії, що дозволить досліджувати можливі варіанти розвитку подій, виявляти приховані взаємозв’язки між концептами та підвищувати ефективність заходів протидії.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Hwang, T. (2021). Deepfakes - Primer and Forecast. Riga: NATO Strategic Communications Centre of Excellence.

2. Giles K., Hartmann K., Mustaffa M. (2019). The Role of Deepfakes in Malign Influence Campaigns. Riga: NATO Strategic Communications Centre of Excellence.

3. Юртаєва К.В. Кримінологічний аналіз використання технології Deepfake: коли фейк стає злочином. Вісник кримінологічної асоціації України. 2021. № 1(24). С. 31–42.

4. Вальорска М. А. Діпфейк та дезінформація : практ. посіб. / Аґнєшка М. Вальорска ; пер. з нім. В. Олійника Київ : Академія української преси ; Центр Вільної Преси, 2020. 36 с.

5. Cinar, Burak. (2023). Deepfakes in Cyber Warfare: Threats, Detection, Techniques and Countermeasures. Asian Journal of Research in Computer Science. 16. 178-193. 10.9734/ajrcos/2023/v16i4381 – URL: https://doi.org/10.9734/ajrcos/2023/v16i4381 .

6. Lee S.-F., Fung B. C. M. Deep Fakes and Big Data: the Next Level of Cyber Warfare. The McGill Data Mining and Security (DMaS) Lab. URL: https://dmas.lab.mcgill.ca/fung/pub/LF21thehilltimes_preprint.pdf .

7. Twomey J. J., Linehan C., Murphy G. Deepfakes in warfare: new concerns emerge from their use around the Russian invasion of Ukraine. The Conversation. 26.10.2023. URL: https://theconversation.com/deepfakes-in-warfare-new-concerns-emerge-from-their-use-around-the-russian-invasion-of-ukraine-216393.

8. Noorma M., Demediuk S., Dubynskyi G. A Decade in the Trenches of Cyberwarfare: Ukraine’s Story of Resilience. Kyiv International Cyber Resilience Forum. 05.02.2024. URL: https://cyberforumkyiv.org/A_Decade_in_the_Trenches_of_Cyberwarfare.pdf .

9. Dilrukshi Gamage, Piyush Ghasiya, Vamshi Bonagiri, Mark E. Whiting, and Kazutoshi Sasahara. 2022. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA. ACM, New York, NY, USA 19 Pages. URL: https://doi.org/10.1145/3491102.3517446 .

10. Sample I. What are deepfakes – and how can you spot them? [Електронний ресурс] / Ian Sample. – 2020. – URL: https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them .

11. Самчинська О. А. Інформаційне насильство, інформаційна маніпуляція та пропаганда: поняття, ознаки та співвідношення / О. А. Самчинська, В. М. Фурашев // Інформація і право. – 2021. – № 1. –С. 55-65. – URL: http://nbuv.gov.ua/UJRN/Infpr_2021_1_8 .

12. Основи кібербезпеки та кібероборони: підручник / Ю.Г. Даник, П.П. Воробієнко, В.М. Чернега. – [Видання друге, перероб. та доп.]. – Одеса: ОНАЗ ім. О.С.Попова, 2019. – 320 с. – URL: https://metod.onat.edu.ua/download/686 .

13. Nair R. What are the Effects of Cyberwarfare? [Електронний ресурс] / Revathy Nair. – 2022. – URL: https://www.tutorialspoint.com/what-are-the-effects-of-cyberwarfare.

14. Martins, Sofia. (2020). The Dark Side of Interconnectivity: Social Media as a Cyber–Weapon?. 10.1007/978-3-030-47511-6_11 – URL: https://www.researchgate.net/publication/344263639_The_Dark_Side_of_Interconnectivity_Social_Media_as_a_Cyber-Weapon .

15. Goodfellow, Ian & Pouget-Abadie, Jean & Mirza, Mehdi & Xu, Bing & Warde-Farley, David & Ozair, Sherjil & Courville, Aaron & Bengio, Y.. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3. 10.1145/3422622 – URL: https://www.researchgate.net/publication/263012109_Generative_Adversarial_Networks .

16. Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., ... & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, 103525. URL: https://doi.org/10.1016/j.cviu.2022.103525 .

17. Difference between War and Warfare. Stack Exchange - English Language Usage. 06.11.2020. URL: https://english.stackexchange.com/questions/550987/difference-between-war-and-warfare .

18. Cyberspace Operations Concept Capability Plan 2016-2028. National Technical Reports Library. 22.02.2010.

19. Даник Ю. Г., Вдовенко С. Г. Концептуальні напрями комплексного вирішення проблеми захисту інформації в системі скритого управління збройних сил. Сучасні інформаційні технології у сфері безпеки та оборони, 2017. № 2(29). С. 98–107.

20. Boyte, Kenneth. (2019). The Evolution of Cyber Warfare in Information Operations Targeting Estonia, the U.S., and Ukraine. URL: http://dx.doi.org/10.4018/978-1-5225-8304-2.ch007 .

21. Buxton O. Cyber Warfare: Types, Examples, and How to Stay Safe. Avast. Avast Academy : Security. 14.07.2023. URL: https://www.avast.com/c-cyber-warfare .

22. Al-Durrah, Qusai & Sadkhan, Sattar. (2021). Cyberwarfare Techniques: Status, Challenges and Future trends. 124-129. URL: http://dx.doi.org/10.1109/ICCITM53167.2021.9677861 .

23. Slonopas A. What Is Cyber Warfare? Various Strategies for Preventing It. American Public University. Information Technology Resources. 16.04.2024. URL: https://www.apu.apus.edu/area-of-study/information-technology/resources/what-is-cyber-warfare/ .

24. Gillis A. Definition Cyberwarfare. TechTarget. Threats and vulnerabilities. 31.03.2023. URL: https://www.techtarget.com/searchsecurity/definition/cyberwarfare .

25. International humanitarian law and policy on Cyber and information operations. International Committee of the Red Cross. Cyber and information operations. URL: https://www.icrc.org/en/law-and-policy/cyber-and-information-operations .

26. DeSombre, W., Campobasso, M., Allodi, L., Shires, J., Work, JD., Morgus, R., O'Neill, P. H., & Herr, T. (2021). A primer on the proliferation of offensive cyber capabilities. In-Depth Research & Reports : Issue Brief. URL: https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/a-primer-on-the-proliferation-of-offensivecyber-capabilities/ .

27. Cyber Warfare. Imperva a Thales Company. Application Security. URL: https://www.imperva.com/learn/application-security/cyber-warfare/ .

28. Mueller, G. B., Jensen, B., Valeriano, B., Maness, R. C., & Macias, J. M. (2023). Cyber Operations during the Russo-Ukrainian War: From Strange Patterns to Alternative Futures. Center for Strategic and International Studies (CSIS). URL: http://www.jstor.org/stable/resrep52130 .

29. Trajcheva S. Social Bots: How Do They Shape Public Opinion?. CHEQ. Cyber Risks & Threats. 13.07.2023. URL: https://cheq.ai/blog/social-bots-how-do-they-shape-public-opinion/ .

30. Kosko B. Fuzzy Cognitive Maps. International Journal of Man-Machine Studies. 1986. Vol. 24(1). P. 65–75. URL: https://doi.org/10.1016/S0020-7373(86)80040-2 .

31. Mental Modeler [Електронний ресурс] – URL: https://www.mentalmodeler.com .

32. DiResta, Renée & Goldstein, Josh. (2024). How spammers and scammers leverage AI-generated images on Facebook for audience growth. Harvard Kennedy School Misinformation Review. URL: http://dx.doi.org/10.37016/mr-2020-151 .

33. Як ШІ-згенеровані зображення стають інструментом маніпуляції?. Facebook. Stopfake. 18.12.2024. URL: https://www.facebook.com/stopfakeukraine/posts/pfbid02DiUfv1QTQR8fPHcJsPYgZ4cCuHhKnBcRSQsFDWnLfGuR5hSfuXYEorkdp2E4Vygnl.

34. Даньшина К. Що трапилось з Facebook? Соцмережі Meta потонули у фейках, а згенерованими позначають реальні фото. IT Community. 26.06.2024. URL: https://itc.ua/ua/novini/shho-trapylos-z-facebook-sotsmerezhi-meta-potonuly-u-fejkah-a-zgenerovanymy-poznachayut-realni-foto/.

35. Небезпека ші-зображень з нібито українськими військовими у соцмережах. Facebook. Центр протидії дезінформації. 26.09.2024. URL: https://www.facebook.com/protydiyadezinformatsiyi.cpd/posts/pfbid0ZGkMg6xFhzFdi5p1EiSQoJkaoAGZGXqPGbVSiP5Wn7oWWmSBxTwwx8PszJ3aJ3Qul.

36. Поширення у Facebook фото з українськими солдатами, які згенерував ШІ. Telegram. Так люблю той Львів. 19.01.2025. URL: https://t.me/Lv1256/72956.

37. Can consumers spot AI and real photos in 2024?. Conjointly. Blog. 17.10.2024. URL: https://conjointly.com/blog/generative-ai-survey-2024/.

38. Amos Z. How Good Are People at Detecting AI?. Unite.ai. Artificial Intelligence. 27.11.2024. URL: https://www.unite.ai/how-good-are-people-at-detecting-ai/.

39. 76% of US consumers unable to spot AI-generated images in new test. Wise Up PR. Insights. 16.05.2024. URL: https://insights.wiseup.pr/76-of-us-consumers-unable-to-spot-ai-generated-images-in-new-test/.

40. AI vs. Human study: Can consumers tell the difference between AI and human-generated content?. Nexcess. Resources. 14.06.2023. URL: https://www.nexcess.net/resources/ai-vs-human-study.

41. Pocol, A., Istead, L., Siu, S., Mokhtari, S., Kodeiri, S. (2024). Seeing is No Longer Believing: A Survey on the State of Deepfakes, AI-Generated Humans, and Other Nonveridical Media. In: Sheng, B., Bi, L., Kim, J., Magnenat-Thalmann, N., Thalmann, D. (eds) Advances in Computer Graphics. CGI 2023. Lecture Notes in Computer Science, vol 14496. Springer, Cham. URL: https://doi.org/10.1007/978-3-031-50072-5_34 .

42. Головне управління розвідки МО України. Обережно, ще одна провокація РФ! [Електронний ресурс] / Головне управління розвідки МО України. – 2022. – URL: https://youtu.be/titiYg843Kk .

43. Wakefield J. Deepfake presidents used in Russia-Ukraine war [Електронний ресурс] / Jane Wakefield. – 2022. – URL: https://www.bbc.com/news/technology-60780142 .

44. Walter, N., & Tukachinsky, R. (2019). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? Communication Research. URL: https://doi.org/10.1177/0093650219854600

ДОДАТОК А

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

Тези наукових доповідей:

1. Прищепа М. О., Наконечний В. С. Модель впливу комплексної протидії технології Deepfake. Проблеми кібербезпеки інформаційно-комунікаційних систем: Збірник матеріалів доповідей та тез; м. Київ, 11 квітня 2025 року; Київський національний університет імені Тараса Шевченка / Редкол.: В.В. Ільченко, д.ф-м.н., проф., (голова); та ін. – К.: ВПЦ "Київський університет", 2025. – с. 10–12.