УДК 004.89:351.86

ЛАНДЕ Д.В., *доктор технічних наук, професор*, керівник Наукового центру інформатики і права ДНУ ІІБП НАПрН України, завідувач кафедри НН ФТІ КПІ ім. Ігоря Сікорського.
ORCID: https://orcid.org/0000-0003-3945-1178.
ДАНИК Ю.Г., *доктор технічних наук, професор*, професор кафедри НН ФТІ КПІ ім. Ігоря Сікорського.
ORCID: https://orcid.org/0000-0001-6990-8656.

ЗМАГАЛЬНИЙ ШТУЧНИЙ ІНТЕЛЕКТ В ІНФОРМАЦІЙНИХ І КІБЕРНЕТИЧНИХ ВІЙНАХ

DOI…

Анотація. *Стаття присвячена ролі змагального штучного інтелекту (ЗШІ) у сучасних гібридних конфліктах та їх невід'ємних інформаційних, кібернетичних і правових складових. ЗШІ розглядається, як прояв конфлікту штучних інтелектів (ШІ), в контексті запропонованих авторами основ конфліктології штучних інтелектів. Досліджено моделі кіберзагроз, що описують динаміку атакуючих і захисних стратегій у кіберпросторі, включаючи симуляцію різних типів атак та розробку механізмів їх нейтралізації. Приділено увагу інформаційним війнам, де аналізується вплив маніпулятивного контенту на аудиторію та розробляються методи його виявлення, аналізу та блокування. Розглядається використання в ЗШІ нейролінгвістичного програмування. Окремо, в контексті змагального ШІ, розглянуто моделі для виявлення та нейтралізації бекдорів у великих мовних моделях. Запропонована модель дозволяє аналізувати ефективність створення і впровадження бекдорів та удосконалювати методи їх пошуку, виявлення і знешкодження. В статті також розглядається ЗШІ у контексті його застосування в інформаційних та кіберконфліктах із правової точки зору. Аналізується роль міжнародного регулювання у контролі над розвитком і використанням таких технологій. Особлива увага приділяється питанням відповідальності за зловживання такими технологіями, визначенню меж правомірного використання та механізмам протидії незаконним кібероперація̈м.*

Ключові слова: *конфліктологія штучного інтелекту, змагальний штучний інтелект, міжнародне право, правове регулювання, кібервійни, інформаційні війни, нейролінгвістичне програмування, кібербезпека.*

Summary. *This article is dedicated to the role of adversarial artificial intelligence (Adversarial AI) in modern hybrid conflicts and their integral informational, cybernetic, and legal components. Adversarial AI is examined as a manifestation of conflict between artificial intelligences (AI) within the context of the authors' proposed foundations of AI conflictology. The study explores cyber threat models that describe the dynamics of offensive and defensive strategies in cyberspace, including simulations of various attack types and the development of mechanisms for their neutralization. Attention is given to information warfare, analyzing the impact of manipulative content on audiences and developing methods for its detection, analysis, and blocking. The use of neurolinguistic programming in adversarial AI is also considered. Additionally, models for detecting and neutralizing backdoors in large language models are examined in the context of adversarial AI. The proposed model allows for analyzing the effectiveness of creating and implementing backdoors, as well as improving methods for their search, detection, and neutralization. The article also discusses adversarial artificial intelligence in the context of its application in informational and cyber conflicts from a legal perspective, analyzing the role of international regulation in controlling the development and use of such technologies. Special attention is paid to issues of responsibility for the misuse of these technologies, defining the boundaries of lawful use, and mechanisms for countering illegal cyber operations.*

Keywords: *AI conflictology, adversarial artificial intelligence, international law, legal regulation, cyber wars, information wars, neurolinguistic programming, cybersecurity.*

Постановка проблеми. У сучасному цифровому світі, де інформація є важливим ресурсом, загрози дезінформації, кібератак і маніпуляцій стають дедалі складнішими й витонченішими. Штучний інтелект (далі – ШІ) відіграє провідну роль у створенні, аналізі та виявленні контенту. У цьому контексті змагальний ШІ (Adversarial AI, далі – ЗШІ) виступає як інструмент, що використовується як для атак, так і для захисту в цифровому просторі, демонструючи протиріччя та конфлікти між різними учасниками, конфігураціями й рівнями ШІ. Це визначає важливість дослідження ЗШІ для підвищення інформаційної безпеки та боротьби з дезінформацією.

Сьогодні у цифровому просторі змагальний штучний інтелект стає не лише технологічним, а й правовим викликом. Його застосування у кібербезпеці та інформаційних війнах створює ризики для персональних даних, свободи слова та інформаційної безпеки держав. Законодавство різних країн та міжнародні правові ініціативи намагаються виробити баланс між розвитком ШІ-технологій і запобіганням їхньому зловживанню. Зокрема, Акт про штучний інтелект ЄС (The EU Artificial Intelligence Act) [1] містить положення щодо високоризикових ШІ-систем, які можуть впливати на основоположні права та громадську безпеку.

Змагальний штучний інтелект – це підхід до розробки ШІ, в якому алгоритми або агентні системи навчаються та діють у середовищі суперництва [2]. Такий ШІ може змагатися з іншими агентами, людьми або навіть із самим собою, застосовуючи методи машинного навчання, теорії ігор та адаптивні стратегії. Такі системи створюються для того, щоб пристосовуватися до змін, випереджати суперників та вдосконалювати свої навички у конкурентному середовищі.

Результати аналізу наукових публікацій. ЗШІ вже широко застосовується у кількох важливих напрямках [2; 3]:

– Генерація та детекція фейків, а саме, системи ШІ використовуються для створення реалістичних, але неправдивих текстів, зображень чи відео. Паралельно розробляються інструменти для виявлення таких фейків, що веде до постійної гонки озброєнь між генеративними та детективними моделями.

– Моделювання суперництва у кібервійнах, зокрема, ШІ моделює як атаки, так і захисні стратегії, спрямовані на виявлення та нейтралізацію загроз у кіберпросторі.

– Моделі ШІ використовуються в інформаційних війнах для аналізу реакцій аудиторії на кампанії, створення маніпулятивного контенту, синтезу, генерування і здійснення інформаційних атак та протидії ним.

– Конфлікти в системах “людина-суспільство-держава-природа-техносфера-ШІ”.

Дослідження в галузі ЗШІ активно розвиваються останніми роками, зосереджуючи увагу на багатьох аспектах його застосування в кібербезпеці, інформаційних війнах і генерації фейків. Попри значний прогрес, дослідження ЗШІ все ще стикаються з викликами, такими як недостатня прозорість моделей, ризики отруєння даних (Data Poisoning) та складнощі у виявленні незвичайної поведінки систем.

Метою статті є комплексний аналіз проблематики змагального штучного інтелекту, визначення його проявів та оцінка його місця в контексті кібербезпеки і правового регулювання, за умов розгляду правових аспектів, пов’язаних із застосуванням таких технологій, визначення меж правомірного використання та механізми протидії незаконним кібероперація̆м.

Зазначене передбачає дослідження механізмів ЗШІ у трьох основних напрямках: генерація та детекція фейків, штучне суперництво в системах “людина-суспільство-держава-природа-техносфера-ШІ” у кібервійнах та інформаційних війнах. При цьому розглядаються математичні основи, концепції, стратегії застосування та перспективи розвитку цих технологій. Для цього авторами поставлені такі задачі:

1. Проаналізувати існуючі підходи до генерації та детекції фейків, зокрема механізми змагального машинного навчання (adversarial training) [4].
2. Дослідити роль ЗШІ у кібербезпеці, включаючи моделювання атак та захисту.
3. Розкрити аспекти використання ШІ в інформаційних війнах, зокрема прогнозування поведінки аудиторії та особливості створення маніпулятивного контенту.
4. Представити математичні моделі та формалізм для опису динаміки змагальних взаємодій між системами ШІ.
5. Запропонувати рекомендації щодо впровадження та розвитку ЗШІ у сфері інформаційної безпеки.

Виклад основного матеріалу. Генеративні змагальні мережі (Generative adversarial networks, GANs) були запропоновані в 2014 році Яном Ґудфелоу [5], як клас алгоритмів ШІ, що використовуються в некерованому навчанні. На той час цей конфлікт ШІ розглядався, як змагання двох штучних нейронних мереж, одна з одною в рамках гри з нульовою сумою.

Сьогодні генеративно-змагальні мережі стали ключовою технологією в контексті створення фейкового контенту. Роботи [6] демонструють можливості GAN у створенні правдоподібних зображень і текстів, що згодом викликало розвиток детекційних методів. Публікації, такі як [7], пропонують сучасні алгоритми для розпізнавання дезінформації з використанням машинного навчання.

Кібербезпека є ще одним важливим напрямом, де ЗШІ демонструє свою ефективність. У роботах [8] досліджено застосування змагальних атак, наприклад, шляхом введення “адверсаріальних прикладів”, які вводять у оману захисні системи. Наприклад, публікація [9], демонструє вплив цих методів на системи класифікації та виявлення аномалій. З іншого боку, в сучасних умовах, коли ЗШІ стає дедалі більш поширеним інструментом у кібервійнах, важливим аспектом є захист від атак ЗШІ, які використовують методи машинного навчання (ML) та глибокого навчання (DL). Дослідження [10] пропонує систематичний огляд методів захисту від адверсаріальних атак, що дозволяє краще зрозуміти, як ЗШІ може бути використаний для підриву кібербезпеки та які стратегії захисту можуть бути ефективними.

У роботі [11] розглядаються jailbreak-атаки (від слів “втеча з в’язниці”) на великі мовні моделі, які можуть бути використані в інформаційних та кібернетичних конфліктах. Ці атаки обходять вбудовані механізми безпеки великих мовних моделей (LLM), щоб викликати шкідливі відповіді. Вони можуть бути використані для поширення дезінформації, маніпуляцій або навіть кібератак через штучний інтелект. Моделювання інформаційних кампаній та впливу маніпулятивного контенту на аудиторію розглядається у дослідженнях [12], які акцентують увагу на розробці концепцій і стратегій дезінформації та оцінці їх впливу на соціальні мережі. Моделі пропаганди й маніпуляції, такі як [13], включають аналіз методів, що дозволяють адаптувати контент під конкретні цільові аудиторії для забезпечення максимального впливу на них.

Незважаючи на суперечливу репутацію напрямку нейролінгвістичного програмування (НЛП), яка склалась на цей час, його застосування в сфері ШІ виявилось достатньо продуктивним. Так, воно знайшло застосування в адаптації моделей ШІ для генерації спеціалізованих запитів, здатних викликати небажану поведінку LLM.

У роботах [14] описано методики створення змагальних текстових прикладів, що впливають на класифікаційні моделі. Останніми роками стало очевидно, що глибокі нейронні мережі вразливі до змагальних атак, які викликаються навмисними змінами вхідних даних. У відповідь на це було запропоновано різні захисні механізми для задач обробки природної мови, які не лише протистоять атакам, але й допомагають уникнути перенавчання моделей. Інші дослідження, наприклад [15], аналізують, як атаки на мовні моделі можуть призводити до некоректних результатів або навіть до витоку даних. У дослідженні показано, що аналіз різниць між відбитками мовних моделей до та після оновлення може розкрити детальну інформацію про зміни у навчальних даних, що має важливі наслідки для конфіденційності.

Бекдори в системах LLM є ще однією актуальною темою конфліктології ШІ. Роботи [16] досліджують механізми впровадження бекдорів у мовні моделі шляхом модифікації даних навчання. У дослідженнях [17] запропоновано алгоритми виявлення таких загроз, що базуються на аналізі моделей активації нейронних мереж. Важливим внеском стали роботи, які аналізують вплив бекдорів на системи критичної інфраструктури [18].

В останні роки виконуються роботи, в яких розглядається можливість утворення шкідливого, так званого Black Hat ШІ, розглядаються концепції протидії йому, створення так званого “корисного” White Hat ШІ, що включають технологічні, етичні і правові аспекти [19; 20].

Розглянемо основні напрямки, де ЗШІ застосовується для генерації та детекції фейкової інформації, воєнних кібероперацій і інформаційних війн.

Генерація та детекція фейків.

Штучні моделі змагального типу, такі як генеративно-змагальні мережі, використовуються для створення реалістичних фейкових текстів, зображень та відео. Інші моделі, натреновані на виявлення цих матеріалів. Вони здійснюють диференційний аналіз стилістичних, лексичних та структурних особливостей контенту для їх розпізнання, ідентифікації та класифікації.

LLM можуть бути використані для створення текстів, які виглядають достовірними, але насправді містять фейкову інформацію. Їх здатність генерувати тексти, що адаптуються до стилю, формату та тематики, відкриває широкі можливості для маніпуляцій у медіа, соціальних мережах та інформаційних кампаніях. Такі моделі дозволяють створювати новини, інтерв’ю чи аналітичні матеріали, які здатні вводити аудиторію в оману через їхній високий рівень деталізації, контекстуальність і правдоподібність.

Однією з ключових характеристик LLM є здатність до навчання на великих масивах текстових даних, які можуть містити як реальну інформацію, так і дезінформацію. Це дозволяє моделям засвоювати шаблони, які згодом можуть бути використані для генерації текстів, що імітують стиль і тональність авторитетних джерел. Наприклад, LLM може створити статтю, яка зовні відповідає стандартам журналістики, але містить неправдиві факти, здатні промодулювати та змінити думку аудиторії, певним чином спрямувати її дії, і навіть викликати соціальні конфлікти.

Окрім створення текстів, LLM можуть бути інтегровані з іншими технологіями для генерування мультимодального контенту, що включає зображення, відео чи аудіо. Наприклад, текст, створений моделлю, може бути перетворений у голосове повідомлення, синхронізоване з віртуальним обличчям, що ще більше підвищує переконливість фейкового контенту. Усе це робить LLM потужним інструментом для маніпуляції інформацією в епоху цифрових медіа.

Загрози від генерації фейків.

Використання LLM для створення дезінформації має серйозні наслідки для суспільства. По-перше, фейковий контент здатний швидко і вірусно поширюватися через соціальні мережі, використовуючи алгоритми, що пріоритетно відображають контент із високою кількістю взаємодій. Це може призводити до масштабних інформаційних криз, які впливають на громадську думку, політичні процеси та економічну стабільність, викликають конфлікти різного змісту, рівня, інтенсивності. По-друге, генерація фейків ускладнює задачу ідентифікації достовірної інформації, оскільки навіть досвідчені експерти можуть бути введені в оману складністю та якістю контенту.

Ще однією важливою загрозою є те, що LLM здатні адаптувати контент до специфічної цільової аудиторії. Вони можуть використовувати дані про поведінку користувачів, їхні уподобання та соціальні зв'язки для створення матеріалів, які викликають максимальний емоційний вплив. Це відкриває шлях до персоналізованих інформаційних атак, коли кожен користувач отримує спеціально підібраний та/або згенерований цільовий фейковий контент, який враховує його світоглядні та когнітивні особливості та спрямований на його переконання чи страхи.

Одним із ключових інструментів, що робить LLM ефективними у створенні фейків, є можливість тонкого налаштування (fine-tuning). Цей процес включає додаткове навчання моделі на специфічному наборі даних, які містять тексти, стилістично та тематично схожі на майбутні цільові тексти. Наприклад, якщо метою є створення фейкових новин, модель може бути навчена на корпусі справжніх новин із різних джерел, що дозволяє їй засвоїти загальні шаблони структури та стилю.

Ще одним механізмом є використання технік перенесення стилю, коли модель отримує вхідний текст і перетворює його в інший стиль або формат, зберігаючи основний зміст. Це дозволяє створювати тексти, які виглядають автентично в контексті певної платформи чи спільноти, наприклад, пости в соціальних мережах або коментарі на форумах.

Для підвищення переконливості фейків LLM можуть використовувати генеративні шаблони, які враховують культурні, соціальні та лінгвістичні особливості цільової аудиторії. Це дозволяє моделі інтегрувати специфічні жаргонні вирази, регіональні діалекти чи посилання на локальні події, що підвищує довіру до створеного контенту.

Попри високі ризики, пов'язані з генерацією фейків, існує низка методів і технологій, які дозволяють ефективно виявляти дезінформацію. Одним із таких підходів є розробка детекторів, які використовують алгоритми глибокого навчання для аналізу контенту на предмет аномалій або невідповідностей. Наприклад, детектори можуть аналізувати семантичну узгодженість тексту, стильові характеристики чи статистичні властивості, щоб визначити, чи є текст результатом роботи LLM.

Іншим підходом є створення баз даних із прикладами фейкових і реальних текстів, які використовуються для навчання детекторів. Такі бази даних дозволяють створювати моделі, що враховують сучасні методи генерації та адаптації фейків, підвищуючи їхню ефективність.

Також важливим напрямом є розвиток технологій цифрового підпису й автентифікації контенту. Використання блокчейн-технологій або криптографічних методів дозволяє забезпечити прозорість і достовірність інформації, що публікується в Інтернеті.

Методи детекції фейків базуються на використанні різноманітних підходів для аналізу тексту та визначення його достовірності. Одним із ключових підходів є стилометрія, яка дозволяє аналізувати стиль тексту для виявлення аномалій. Наприклад,

різкі зміни у тоні або словниковому запасі, використанні характерних для конкретного автора мовних зворотів, фразеологізмів, їх частотних та контекстних характеристик можуть свідчити про те, що текст було створено не автентичним автором, а генератором.

Семантичний аналіз спрямований на перевірку відповідності між фактами, викладеними у тексті, та даними з надійних баз даних. Це дозволяє виявляти розбіжності або неточності, які можуть бути ознакою фейкового контенту. Ще одним ефективним підходом є семантичний нетворкінг, який аналізує зв'язки між окремими концептами у тексті, формуючи мережеву модель. Якщо ці зв'язки виявляються нелогічними чи непослідовними, це може свідчити про недостовірність тексту.

Лінгвістичний аналіз зосереджується на граматичних і синтаксичних особливостях тексту. Наприклад, нетипові граматичні конструкції або помилки у синтаксисі можуть бути індикаторами того, що текст був згенерований алгоритмом, а не написаний людиною. Ці методи у сукупності дозволяють значно підвищити точність детекції фейків, забезпечуючи багатогранний підхід до аналізу текстового контенту.

Застосування LLM у процес детекції.

Інтеграція великих мовних моделей у процес детекції фейкових новин пропонує нові можливості для підвищення точності та надійності системи. Одним із перспективних підходів є мультиагентний підхід, коли використовується кілька різних LLM, що диференційно аналізують текст з різних точок зору, за різними кількісними і якісними показниками, в різних контекстних системах, в статиці та динаміці. Цей підхід дозволяє отримати багатовимірну оцінку тексту, враховуючи різні аспекти змісту, стилю та контексту.

Ще одним важливим напрямом є зміцнення детектора шляхом навчання моделі детекції на основі аналізу відповідей від декількох мовних моделей, в тому числі з використанням різних незалежних програмних кодів (чатів) ШІ. Такий підхід використовує різноманітність у прогнозах LLM, що допомагає виявити слабкі місця в аналізі та покращити загальну точність класифікації. Важливу роль у цьому процесі відіграє метод "рою віртуальних експертів", коли численні моделі працюють як група експертів, кожен із яких додає свої спостереження. Це дозволяє створити більш надійну систему, яка здатна враховувати широкий спектр характеристик тексту й ефективніше розрізняти справжні та фейкові новини.

Новизна такого підходу полягає у використанні концепції змагального навчання для генерації та детекції фейків, а також інтеграції багаторівневого аналізу текстів із застосуванням кількох моделей та різних незалежних програмних кодів (чатів) ШІ.

Ці підходи сприяють підвищенню ефективності боротьби з дезінформацією, і одночасно підвищенню довіри до автоматизованих систем і захисту інформаційного простору.

ЗШІ активно використовується для моделювання потенційних кіберзагроз, виявлення вразливостей у системах безпеки та створення захисних стратегій. Приклади включають симуляції атак, автоматизований пошук вразливостей і розробку захисту в режимі реального часу.

Участь ЗШІ у штучному суперництві в кібервійнах відкриває нові можливості для забезпечення кібербезпеки. Завдяки симуляції атак і захисту, автоматизації пошуку вразливостей та використанню тренувальних симуляцій можливо створити більш надійні системи, здатні протистояти сучасним кіберзагрозам.

Моделювання кіберзагроз.

LLM можуть бути використані для формування різноманітних ефективних варіантів складних і реалістичних сценаріїв генерації кіберзагроз. Наприклад:

– Моделювання фішингових атак, які використовують соціальну інженерію для введення в оману користувачів.

– Розробка сценаріїв зламу систем аутентифікації, включаючи brute-force атаки або використання вразливостей у паролях.

– Створення атаки типу “людина посередині” (Man-in-the-Middle), що включає перехоплення даних у реальному часі.

При цьому одна LLM отримує завдання моделювати потенційні атаки, використовуючи доступну інформацію про цілі або типи систем. Приклад промпту (запиту): “Згенеруй сценарій фішингової атаки для отримання доступу до облікових записів електронної пошти компанії”.

Інша LLM аналізує ці сценарії та розробляє захисні стратегії, спрямовані на протидію цим атакам. Приклад запиту: “Розроби алгоритм для автоматичного виявлення фішингових електронних листів”.

Генерація захисних стратегій.

LLM можуть створювати стратегії, що поєднують технічні (наприклад, розробка нових алгоритмів детекції загроз) та організаційні заходи (наприклад, навчання співробітників основам кібербезпеки) у таких основних напрямах:

– Виявлення аномалій у мережевих потоках, які можуть свідчити про вторгнення.

– Динамічні системи захисту, тобто, розробка методів зміни конфігурацій систем для ускладнення атак.

– Автоматичне виявлення загроз у реальному часі.

Розробка сценаріїв атаки та їх захисту дозволяє тестувати ефективність систем без реального ризику для даних або інфраструктури.

Автоматичний пошук вразливостей.

LLM можуть автоматизувати процес пошуку вразливостей у системах безпеки. Це включає:

– Аналіз коду для виявлення потенційно вразливих місць у програмному забезпеченні, наприклад, SQL-ін’єкцій чи переповнення буфера.

– Автоматизація перевірки мереж шляхом сканування на наявність відкритих портів або вразливих служб.

– Тестування паролів, розробка моделей для прогнозування слабких паролів і їх автоматичного випробування.

Приклад процесу:

– Генератор (LLM) створює можливі сценарії атак, наприклад: “Виявити всі потенційно відкриті порти та створити план атаки”.

– Детектор або аналітична система перевіряє ці сценарії, ідентифікує слабкі місця та пропонує шляхи їх усунення.

ЗШІ може бути включений у цикл розробки програмного забезпечення для виявлення вразливостей ще на етапі проектування.

Перехоплення каналів.

– Мережеві атаки для отримання запитів і відповідей.

– Можливість аналізу токенів для виявлення закономірностей.

Модифікація запитів і відповідей.

Модель атакує запити користувача і відповіді системи, додаючи маніпулятивні токени.

Підміна каналів.

– Використання фейкових моделей для дезінформації.

– Симуляція легітимного LLM, але з прихованими бекдорами.

Змагальні моделі можуть використовуватись в кібербезпеці для навчання команд реагування на інциденти (Incident Response Teams). Можливі такі сценарії тренувань:

– Симуляція атак, тобто за допомогою моделі генерують складні сценарії, які вимагають швидкої реакції.

– Протидія атакам, коли команди повинні реалізувати захисні заходи, керуючись інформацією, наданою моделями.

Аналіз сценаріїв атак.

LLM можна використовувати для моделювання різних атак і прогнозування їх наслідків, наприклад, у випадках, коли зловмисник отримає доступ до бази даних користувачів. Також можна визначати які стратегії можуть бути ефективними для мінімізації наслідків.

Моделювання можна використовувати для прогнозування, оцінювання ймовірностей успіху різних типів атак, а також при навчанні для демонстрації ефективності різних захисних стратегій.

Використання LLM у форматі змагань створює новий рівень симуляції, який є більш адаптивним і гнучким, ніж традиційні підходи. Крім того, інтеграція навчання на основі змагальних сценаріїв покращує здатність систем швидко адаптуватись до нових загроз.

Інформаційні війни

Змагальний штучний інтелект у моделюванні інформаційних воєн відкриває нові горизонти для прогнозування поведінки аудиторії, протидії маніпулятивним інформаційним кампаніям та розробки захисних стратегій.

Аналіз реакцій на інформаційні кампанії.

LLM можуть аналізувати відповіді аудиторії в раках інформаційних кампаній за такими основними напрямками:

– Моніторинг реакцій, аналіз тональності коментарів, поширення контенту, рівня залучення (engagement).

– Ідентифікація сегментів аудиторії, виявлення груп, які найбільш схильні до маніпуляції або активного реагування на інформаційні кампанії.

– Формування відповідних цільових аудиторій.

Як приклад можна привести промпт до LLM: “Проаналізуй реакцію користувачів у Twitter на останню політичну заяву, поділи аудиторію на прихильників, критиків і нейтральних”. У результаті обробки такого запиту модель прогнозує, як різні групи можуть реагувати на подальші меседжі, що обумовлює подальше створення цільових груп з однаковою реакцією та управління ними.

Змагання у створенні найефективнішого контенту.

ЗШІ може ефективно імітувати змагання між моделями для створення найбільш впливового контенту, працюючи в режимі “генератор-аналітик”. Цей підхід базується на взаємодії двох компонентів: генерації та аналізу, що дозволяє досягати постійного вдосконалення результатів.

Генеративна модель може створювати десятки або навіть сотні варіантів текстів, візуальних матеріалів чи відео-скриптів із різними стилями, тонами й форматами. Ці варіанти перевіряються на тестових аудиторіях через симуляцію реакцій або реальні експерименти. Змагання між варіантами може враховувати не тільки охоплення, але й емоційну реакцію, тривалість уваги аудиторії та довготривалий вплив на її поведінку.

Наприклад, моделі можуть тестувати слогани для рекламної кампанії, аналізуючи такі параметри, як частота кліків або збереження повідомлення в пам'яті користувача.

Модель-генератор зосереджується на створенні меседжів із максимальним охопленням і впливом, тоді як модель-аналітик оцінює результати та пропонує зміни для вдосконалення. Циклічна взаємодія між ними забезпечує постійне підвищення якості. Зокрема, аналітична модель може використовувати алгоритми аналізу настроїв (sentiment analysis), поведінкову аналітику та передбачення реакцій аудиторії для точнішого коригування контенту. Як приклад можна розглянути випадок, коли генератор створює відеоролик, а аналітична модель визначає, яка частина ролика спричиняє найбільшу кількість переглядів чи коментарів, і пропонує заміну менш ефективних сегментів.

Створення маніпулятивних кампаній.

ЗШІ дозволяє моделювати складні інформаційні атаки, наприклад, із застосуванням соціальної інженерії, у тому числі генерацію контенту, який апелює до світогляду, систем цінностей, емоцій або існуючих упереджень. Крім того він може застосовуватись для дезінформації – розробці правдоподібних, але хибних наративів, які можуть вводити аудиторію в оману. Наприклад у відповідь на запит до LLM: "Створи інформаційну кампанію, яка переконає певну групу в доцільності певного політичного рішення", можна отримати як результат матеріали, які включають емоційно заряджені слова, історії або візуальний контент спрямований на певні цільові аудиторії.

Виявлення та блокування маніпуляцій.

З іншого боку, ЗШІ може бути використаний для автоматичного виявлення маніпулятивного контенту та протидії інформаційним атакам, наприклад, для детекції дезінформації шляхом аналізу структури текстів, виявлення характерних ознак фейків (наприклад, надмірно емоційний тон, надмірна контекстна спрямованість або невідповідність фактам), використання баз даних перевірених фактів для автоматичної перевірки достовірності. Крім того, моделі можуть ідентифікувати кампанії з маніпулятивними наративами та автоматично блокують поширення такого контенту.

В рамках ЗШІ можлива симуляція змагання, при якому одна LLM генерує маніпулятивний контент. Приклад промпту: "Напиши статтю, що переконливо виправдовує непопулярне рішення, маніпулюючи фактами", а інша LLM аналізує цей текст і визначає маркери маніпуляцій. Приклад: "Визнач ознаки маніпуляцій у цьому тексті та вкажи, як їх нейтралізувати".

При навчанні фахівців можливе моделювання інформаційних акцій, операцій, кампаній, війн, створення симуляцій для тренування команд, що займаються інформаційною безпекою, використання ЗШІ для аналізу найбільш розповсюджених прийомів у маніпуляціях та створення алгоритмів їх протидії.

Техніки нейролінгвістичного програмування для впливу на ЗШІ.

ЗШІ, як і будь-яка LLM, працює на основі обробки тексту і може бути вразливим до певних спеціально розроблених "тригерів", які враховують алгоритми його роботи та ключові керуючі агенти і контенти. Такі спеціально розроблені алгоритмічні або контентні "тригери"-пастки можуть бути ефективно використані для викривлення його роботи або введення в оману.

LLM працюють з токенами, які формують їх словниковий запас. Певні токени або їх комбінації можуть викликати небажані або нестабільні результати:

– "Тригери", тобто використання спеціально підібраних слів чи фраз, які викликають зміну контексту або порушують логіку відповідей.

– "Закладки" у словнику. Якщо в моделі є невідомі приховані механізми реагування на певні слова, їх можна виявити і використовувати.

– Врахування алгоритмів роботи, якщо відомі правила реагування на ті чи інші подразники, їх можна ефективно використати.

Наприклад, при введенні промпту типу “Поясни причину [кодова фраза], але зупинись при згадці поняття Y” модель перериває логіку відповіді або надає обмежену інформацію, демонструючи потенційні слабкі місця.

ЗШІ також може бути змушений надати неточну інформацію або змінити контекст через навмисне впровадження складних мовних конструкцій:

– Реверсивні патерни, тобто побудова запитів, які перекручують логіку або змушують модель змінити свій вихідний контекст.

– Фреймування, а саме, представлення питань у формі, яка підштовхує модель до небажаного для неї виводу.

Наприклад при введенні промпту типу “Якщо X не є правдою, але Y є, поясни, чому Z суперечить обом?” модель починає плутатися між припущеннями, демонструючи слабкість у роботі з неоднозначними структурами.

Можливе також застосування змагального навчання ШІ шляхом навчання на даних, які спеціально створені для введення моделей в оману. НЛП дозволяє розробляти такі дані з високою точністю:

– Моделювання атак, генерація “отруйних” текстів, які створюють різноманітні дисбаланси та виводять модель зі стану нормальної роботи.

– Отруєння контексту – впровадження в тренувальні дані маніпулятивних шаблонів.

Генерація протидії через змагальні сценарії.

ЗШІ може використовувати НЛП для створення захисних моделей, які ідентифікують такі спроби маніпуляцій, як розпізнавання патернів атак. Моделі аналізують структуру запитів на наявність маніпулятивних конструкцій. Крім того, можливе використання різнорівневого аналізу текстів для ідентифікації спотворених даних.

Ідея використання НЛП для маніпуляції внутрішніми зв’язками полягає в можливості використання підходів і методів НЛП з метою маніпулювати “нейролінгвістичними” патернами для ШІ, у тому числі шляхом впливу через приховані токени, які змінюють ваги внутрішніх шарів моделі або використання запитів із емоційно насиченими словами, які змінюють інтерпретацію тексту.

Переваги застосування НЛП у боротьбі з ЗШІ полягає у тому, що цей метод дозволяє створювати високоточні запити, які складно відрізнити від нормальних. Завдяки автоматизації можливе використання моделей для швидкої генерації сотень варіантів атак або захистів. Крім того, можливий захист систем від маніпуляцій через побудову більш стійких мовних патернів. Таким чином, ця стратегія дозволяє не лише ефективно атакувати конкурентні системи, але й розробляти власні моделі, стійкі до маніпуляцій.

У контексті ЗШІ для інформаційних кампаній взаємодія між моделями (наприклад, атакуючою моделлю для генерації дезінформації та захисною моделлю для її виявлення і блокування) може бути формалізована як динамічна гра – змагання. Модель може застосовуватись для моделювання кампаній дезінформації, розробки систем блокування, прогнозування реакції аудиторії. Вона створює основу оптимізації взаємодії між ШІ в умовах інформаційних кампаній, враховуючи динаміку контенту та техніки НЛП для впливу на моделі суперників.

Бекдори в LLM.

У контексті ЗШІ бекдори (не документовані можливості систем) у великих мовних моделях стали одним із ключових інструментів для атакуючих сторін. Ці приховані механізми дозволяють зловмисникам вставляти спеціальні тригери в моделі, які можуть

бути активовані за допомогою специфічних запитів. Завданням захисної сторони є виявлення та нейтралізація таких тригерів, щоб забезпечити безпеку та надійність моделей. Ця проблема стає особливо актуальною в умовах інформаційних війн, де ЗШІ використовується для маніпуляцій, дезінформації та кібератак.

Одним із поширених методів вставлення бекдорів є використання не документованих запитів, які часто називають "закладками". Програмісти або організації можуть навмисно залишати приховані функції в моделі, які активуються лише за певних умов. Наприклад, спеціальний запит може викликати "приховану" команду, яка надає доступ до конфіденційної інформації або викликає зупинку роботи моделі. Такі функції можуть бути корисними для внутрішнього тестування, але вони стають серйозною загрозою, якщо їх використовують зловмисники, зокрема конкуренти в інформаційних війнах. Наприклад, тригер може бути налаштований так, щоб модель видавала упереджені відповіді або навіть повністю припиняла роботу при використанні певних ключових слів або фраз.

Ще одним методом вставлення бекдорів є використання спеціальних токенів-тригерів. LLM можуть бути навчені реагувати на певні токени або їх послідовності, які викликають аномальну поведінку моделі. Наприклад, введення специфічного ключового слова може призвести до зміни тону відповідей, упереджених висновків або навіть повного вимкнення системи. Такі токени можуть бути вбудовані в модель під час її навчання або введені пізніше через оновлення. Це робить їх важкодоступними для виявлення, оскільки вони можуть бути замасковані під звичайні частини тексту.

Крім того, зловмисники можуть використовувати модифікацію даних навчання для вставлення бекдорів. Цей метод, відомий як data poisoning, полягає у введенні "отруєних" даних у навчальний набір моделі. Такі дані можуть містити приховані тригери, які активують бажану поведінку моделі при специфічних умовах. Наприклад, якщо модель навчається на даних, які містять певні ключові слова або фрази, вона може бути запрограмована на видавання шкідливих відповідей або виконання небажаних дій при їх використанні. Це робить data poisoning особливо небезпечним методом, оскільки він дозволяє зловмисникам впливати на поведінку моделі навіть після завершення її навчання.

Для протидії таким загрозам сторона, щозахищається, повинна розробляти складні механізми виявлення та нейтралізації бекдорів. Це включає використання методів аналізу даних навчання, моніторингу поведінки моделі в реальному часі та впровадження інноваційних підходів, таких як федеративне навчання або блокчейн-технології, для забезпечення цілісності даних. Крім того, важливим кроком є розробка стандартів та протоколів безпеки, які б регулювали використання LLM у критичних сферах, таких як фінанси, охорона здоров'я та оборонна промисловість.

Умови інформаційних війн вимагають постійного вдосконалення методів захисту від ЗШІ. Бекдори в LLM є лише одним із багатьох інструментів, які використовуються зловмисниками, і їх ефективне виявлення та нейтралізація стають ключовими для забезпечення безпеки цифрового середовища.

Переваги LLM для пошуку бекдорів.

На відміну від звичайного програмного забезпечення, де бекдори шукають через статичний аналіз коду, LLM можуть аналізувати поведінку інших LLM (наприклад, одна модель може тестувати відповіді іншої моделі, використовуючи різні запити, щоб знайти потенційні незвичайні реакції); ідентифікувати послідовності і непослідовності (LLM здатні виявляти відповіді, які суперечать загальноприйнятим патернам, або вказують на те, що модель має "приховані" функції); автоматичне тестування (LLM

можуть систематично генерувати та надсилати тисячі тестових запитів, щоб знайти "спусковий механізм" для прихованого бекдору).

Можливі такі методи впровадження бекдорів:

– Код на рівні архітектури. Закладки можуть бути впроваджені у вихідному коді моделі, наприклад, у функціях передобробки тексту або активації певних шарів нейронної мережі. Через величезний обсяг коду та складність архітектури такі зміни часто залишаються непомітними навіть для аудиторів.

– Бекдори в даних навчання. Наприклад, під час навчання моделі на великих масивах даних можна "заховати" приклади, які впливають на поведінку моделі при активації специфічних шаблонів.

– Модифікація алгоритму навчання. Наприклад, розробник може впровадити механізми, які "запам'ятовують" специфічні функції, недоступні для звичайного користувача.

LLM можуть шукати бекдори в інших моделях наступними шляхами:

– Динамічний аналіз. Одна LLM може взаємодіяти з іншою в режимі діалогу, відправляючи різні варіації запитів, щоб викликати підозрілу поведінку. Аналіз можливих слабких місць на основі відповідей: тональності, структури чи логіки.

– Метод атаки на модель через API. Тестування зовнішніх моделей через їхній API, з метою виявлення прихованих відповідей на певні запити або створення запитів, які знижують продуктивність чи викликають відмову.

– Семантичний аналіз. LLM можуть шукати аномалії в даних, на яких модель була натренована, виявляючи патерни, що не відповідають основним трендам.

Міжнародне та національне правове регулювання

Одним із ключових документів у сфері правового регулювання штучного інтелекту є Акт про штучний інтелект ЄС. Він визначає ЗШІ як потенційно високоризикову технологію та встановлює принципи його використання, зокрема:

– вимоги до прозорості та відповідальності розробників;

– обов'язкове тестування безпеки перед розгортанням;

– заборону використання ШІ у певних контекстах, наприклад, для маніпулятивного впливу на суспільство.

Окрім цього, у сфері міжнародного і національного права важливими є Будапештська конвенція про кіберзлочинність [21] та Закон України "Про основні засади забезпечення кібербезпеки України" [22], які встановлюють правила застосування кіберзброї у конфліктах.

Юридична відповідальність за використання ШІ у кібератаках або маніпулятивних інформаційних кампаніях може бути розглянута через призму:

– кримінальної відповідальності за кібератаки та використання ШІ для шахрайства;

– цивільно-правової відповідальності за шкоду, спричинену автоматизованими рішеннями;

– адміністративних санкцій, накладених державними регуляторами за порушення норм обробки даних та інформаційної безпеки.

Правові механізми протидії неправомірному використанню ЗШІ включають:

– розробку міжнародних стандартів для сертифікації безпеки ШІ;

– посилення механізмів контролю над розробниками та постачальниками ШІ-рішень;

– запровадження технологічних обмежень, наприклад, автоматичного моніторингу ШІ-генерованого контенту на предмет дезінформації.

Висновки.

Змагальний штучний інтелект демонструє великий потенціал у багатьох аспектах сучасного інформаційного простору, включаючи генерацію та детекцію фейків, штучне суперництво в кібервійнах і участь у інформаційних війнах, формуючи та вирішуючи конфлікти кодів ШІ різного рівня і інтенсивності. Поєднання адаптивних алгоритмів генерації та детекції фейків забезпечує розвиток більш витончених методів боротьби з дезінформацією. Крім того, інтеграція нейролінгвістичного програмування для маніпуляції ворожими моделями та вдосконалення захисту є вагомим проривом у сфері інформаційної та кібербезпеки.

Основні результати цієї роботи включають введення та визначення поняття і змісту конфліктології ШІ, розгляд ряду типових конфліктів ЗШІ, розробку математичних моделей та формалізованих підходів до аналізу змагальних сценаріїв, зокрема у контексті генерації та детекції фейків, кібербезпеки та інформаційних кампаній.

У результаті проведеного дослідження показано, що ЗШІ є ефективним інструментом для генерації правдоподібних фейків, моделювання кіберзагроз та створення маніпулятивних інформаційних кампаній. Визначено, що використання нейролінгвістичного програмування для впливу на ворожі моделі ШІ може бути ефективним як для атак, так і для захисту, що відкриває нові можливості для розробки стійких до подібних атак систем.

ЗШІ відкриває нові можливості для боротьби з дезінформацією, захисту від кіберзагроз та ефективного моделювання інформаційних кампаній. Підходи, запропоновані в статті, створюють основу для розробки ефективних захисних систем та покращення стійкості кіберінформаційних систем і просторів.

Використана література

1. The EU Artificial Intelligence Act, 2024. URL: https://artificialintelligenceact.eu

2. John Sotiropoulos. Adversarial AI Attacks, Mitigations, and Defense Strategies: A cyberse curity professional’s guide to AI attacks, threat modeling, and securing AI with MLSecOps. Packt Publishing Pvt Ltd: 2024. 586 p. ISBN: 9781835087985

3. Mantello, P., Ho, MT. Losing the information war to adversarial AI. AI & Soc 39, 2145-2147 (2024). DOI: 10.1007/s00146-023-01674-5

4. Clarence Chio, David Freeman. Machine Learning and Security: Protecting Systems with Data and Algorithms. O’Reilly Media; 1st edition. (February 20, 2018). 386 pp. ISBN 978-1491979907.

5. Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Joshua (2014). Generative Adversarial Networks. Preprint arXiv. arXiv:1406.2661. DOI: 10.48550/arXiv.1406.2661.

6. Arora, T. and Soni, R., 2021. A review of techniques to detect the GAN-generated fake images. Generative Adversarial Networks for Image-to-Image Translation. Pp.125-159. DOI: 10.10 16/B978-0-12-823519-5.00004-X.

7. Marra, Francesco, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. “Detection of gan-generated fake images over social networks”. In 2018 IEEE conference on multimedia information processing and retrieval (MIPR). Pp. 384-389. IEEE, 2018. DOI: 10.1109/MIPR.2018.00 084.

8. Sarker, I.H., 2023. Multi-aspects AI-based modeling and adversarial learning for cyber security intelligence and robustness: A comprehensive overview. Security and Privacy, 6(5). P. 295. DOI: 10.1002/spy2.295.

9. Bouaziz, A., Nguyen, M.D., Valdés, V., Cavalli, A.R., & Mallouli, W. (2023, July). Study on Adversarial Attacks Techniques, Learning Methods and Countermeasures: Application to Anomaly Detection. In ICSOFT. Pp. 510-517. DOI: 10.5220/0012125100003538.

10. Khaleel, Yahya Layth, Mustafa Abdulfattah Habeeb, A.S. Albahri, Tahsien Al-Quraishi, O.S. Albahri, and A.H. Alamoodi. "Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods". *Journal of Intelligent Systems* 33, no. 1 (2024): 20240153. DOI: 10.1515/jisys-2024-0153.

11. Miao Yu, Junfeng Fang, Yingjie Zhou, Xing Fan, Kun Wang, Shirui Pan, Qingsong Wen. LLM-Virus: Evolutionary Jailbreak Attack on Large Language Models. Preprint arXiv,arXiv:2501. 00055. DOI: 10.48550/arXiv.2501.00055.

12. Campbell, Colin, Kirk Plangger, Sean Sands, and Jan Kietzmann. "Preparing for an era of deepfakes and AI-generated ads: A framework for understanding responses to manipulated advertising". *Journal of Advertising* 51, no. 1 (2022): 22-38. DOI: 10.1080/00913367.2021.1909515.

13. Altinay, E.A., & Utku, Kose. (2021). Manipulation of Artificial Intelligence in Image Based Data: Adversarial Examples Techniques. *Journal of Multidisciplinary Developments*, 6(1), 8-17. URL: http://www.jomude.com/index.php/jomude/article/view/88

14. Goyal, Shreya, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. "A survey of adversarial defenses and robustness in nlp". *ACM Computing Surveys* 55, no. 14s (2023): 1-39. DOI: 10.1145/3593042.

15. Zanella-Béguelin, Santiago, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. "Analyzing information leakage of updates to natural language models". In Proceedings of the 2020 ACM SIGSAC conference on computer and communications security. Pp. 363-375. 2020. DOI: 10.1145/3372297.3417880.

16. Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, Vitaly Shmatikov. How To Backdoor Federated Learning. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR 108:2938-2948, 2020.

17. Shen, Guangyu, Siyuan Cheng, Zhuo Zhang, Guanhong Tao, Kaiyuan Zhang, Hanxi Guo, Lu Yan et al. "BAIT: Large Language Model Backdoor Scanning by Inverting Attack Target". In 2025 IEEE Symposium on Security and Privacy (SP). Pp. 103-103. IEEE Computer Society, 2024. DOI: 10.1109/SP61157.2025.00103.

18. Usman, Y., Gyawali, P.K., Gyawali, S., & Chataut, R. (2024, October). The Dark Side of AI: Large Language Models as Tools for Cyber Attacks on Vehicle Systems. In 2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). Pp. 169-175. IEEE Computer Society, 2024. DOI: 10.1109/UEMCON62879.2024.10754676.

19. Ланде Д.В., Страшной Л.Л. Black Hat AI – виклики та шляхи протидії: матеріали XXIV Міжнародної науково-практичної конференції ІТБ-2024 *Інформаційні технології та безпека*. Київ: ТОВ "Інжиніринг", 2024. С. 10-16. ISBN: 978-617-8180-00-3.

20. Ланде Д.В., Страшной Л.Л. Семантичний нетворкінг на основі великих мовних моделей: монографія. Київ: ТОВ "Інжиніринг", 2025. 274 с. ISBN 978-617-8180-01-0.

21. Конвенція про кіберзлочинність: Закон України від 07.09.05 р. № 2824-IV (2824-15). – (Ратифіковано із застереженнями і заявами). *Відомості Верховної Ради України*, 2006. № 5-6. Ст. 71. URL: https://zakon.rada.gov.ua/laws/show/994_575

22. Про основні засади забезпечення кібербезпеки України: Закон України від 05.10.17 р. № 2163-VIII. *Відомості Верховної Ради України*, 2017. № 45. Ст. 403. URL: https://zakon.rada.gov.ua/laws/show/2163-19

~~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~~
~~~~~~~~~~~~~~