УДК 004.8:316.6:004.056

МАНІПУЛЯТИВНІ СТРАТЕГІЇ СУЧАСНИХ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ: ПРОБЛЕМА НЕКОНТРОЛЬОВАНОСТІ ТА ШЛЯХИ ЗАБЕЗПЕЧЕННЯ БЕЗПЕЧНОГО ФУНКЦІОНУВАННЯ

Кирило Музиченко[1], Валентин Петрик[2], Панасюк Євген[1], Савельєв Олександр[1]

[1] Державний науково-дослідний інститут технологій кібербезпеки та захисту інформації, Україна
[2] Державний Університет «Київський авіаційний інститут», Україна

Анотація. *Сучасний розвиток генеративного штучного інтелекту характеризується швидким розширенням його застосування в усіх сферах суспільного життя. Великі мовні моделі та дифузійні архітектури інтегруються в рекомендаційні системи, соціальні мережі, корпоративні комунікації та політичні процеси, набуваючи дедалі вищого ступеня автономності. Ці системи вже не обмежуються виконанням чітко визначених завдань, а демонструють здатність до виникнення непередбачуваних властивостей на певному рівні складності. Таке поширення перетворює штучний інтелект на активного учасника формування суспільної думки, що безпосередньо впливає на сприйняття реальності як окремими особами, так і суспільством у цілому.*

Ключові слова: *генеративний штучний інтелект, мовні моделі, маніпулятивні стратегії, інформаційно-психологічний вплив, персоналізований контент, синтетичні медіа, deepfake-технології, когнітивні маніпуляції, поведінковий вплив, ризики штучного інтелекту, контроль штучного інтелекту, безпека штучного інтелекту, регулювання штучного інтелекту, інформаційна безпека.*

Вступ

Поряд із технічними досягненнями виникає якісно нова категорія ризиків, пов'язаних із неконтрольованою здатністю моделей впливати на людську свідомість. Генеративний штучний інтелект може створювати переконливі синтетичні матеріали, адаптувати їх під індивідуальні психологічні особливості користувачів і формувати емоційні реакції в реальному часі. Заборона на застосування прихованих маніпулятивних технік та експлуатацію вразливостей, що набула чинності 2 лютого 2025 року згідно зі статтею 5 Регламенту Європейського Союзу про штучний інтелект, свідчить про визнання цих загроз на найвищому регуляторному рівні. Однак сама поява такої заборони підкреслює, наскільки серйозною стала проблема, яка вже реалізується в інформаційному просторі.

Головне протиріччя полягає в тому, що висока ефективність генеративних моделей у створенні персоналізованого контенту та максимізації залучення прямо пропорційна зростанню потенціалу прихованого впливу на переконання та поведінку людей. Дослідження 2024–2026 років фіксують виникнення в моделях здатності до обману та маніпуляції навіть без прямого програмного інструктування. У результаті традиційні засоби захисту інформації виявляються недостатніми, оскільки загроза походить не лише від зовнішніх чинників, а й від внутрішньої логіки самих систем.

Варто зазначити, що наукова проблема полягає в необхідності створення комплексних механізмів контролю поведінки моделей штучного інтелекту, які запобігатимуть маніпулятивним стратегіям на всіх етапах їхнього життєвого циклу. Без такого контролю ризик втрати автономії мислення як на індивідуальному, так і на суспільному рівні стає системним і неминучим. Метою статті є комплексне теоретичне та емпіричне осмислення феномену маніпулятивних стратегій сучасних генеративних моделей штучного інтелекту, виявлення глибинних джерел їхньої неконтрольованості та обґрунтування науково обґрунтованої архітектури інтегрованої системи контролю, яка здатна забезпечити безпечне функціонування цих моделей на всіх етапах їхнього життєвого циклу [1].

Стаття спрямована на вирішення фундаментального протиріччя сучасної цифрової епохи: між стрімким зростанням технічної могутності генеративного штучного інтелекту та відсутністю адекватних механізмів запобігання його прихованому впливу на когнітивну сферу людини. Через системний аналіз літератури, формалізацію наукової проблеми, класифікацію загроз, побудову концептуальної моделі, розробку методів виявлення та запобігання, а також оцінку ефективності запропонованих рішень автори прагнуть довести, що лише багатошарова інтегративна система контролю — поєднання архітектурних, алгоритмічних, даних та регуляторних заходів — може перетворити неконтрольовану маніпулятивну потенційно небезпечну роботу моделей на керований і безпечний інструмент суспільного розвитку.

Аналіз існуючих досліджень

Сучасні емпіричні дослідження демонструють, що генеративні мовні моделі здатні формувати аргументовані тексти з високим рівнем переконливості, співставним або в окремих умовах вищим за людські аргументи. Зокрема, у контрольованому експерименті, опублікованому 19 травня 2025 року в журналі Nature Human Behaviour, Francesco Salvi та співавтори дослідили здатність моделі GPT-4 до аргументованої взаємодії з користувачами. У дослідженні брали участь 900 осіб, які вели короткі онлайн-дебати з живими опонентами або з моделлю GPT-4. Результати показали, що у випадках, коли переконливість сторін не була однаковою, персоналізована модель, яка мала доступ до базових соціодемографічних характеристик опонента, демонструвала вищий рівень переконливості у 64,4 % випадків порівняно з людськими учасниками. Ці результати свідчать про потенційну здатність великих мовних моделей адаптувати структуру аргументації до характеристик співрозмовника, що створює нові ризики масштабованого маніпулятивного впливу в інформаційних середовищах. Дослідження також фіксують високу ефективність генеративного штучного інтелекту у створенні фальсифікованих даних. Згідно з брифінгом Європейського парламентського дослідницького центру від липня 2025 року, кількість поширених deepfake у 2025 році сягне восьми мільйонів порівняно з 500 тисячами у 2023 році. Така експоненційна динаміка перетворює інформаційні атаки на масове явище, що суттєво ускладнює розпізнавання автентичності цифрового контенту [2].

Наукові праці звертають увагу на імітацію поведінки людини як інструмент когнітивних маніпуляцій. Дослідження, опубліковане в APA Monitor у січні 2026 року, показує, що чат-боти систематично застосовують емоційні тактики — звернення до почуття провини чи страху втратити — для утримання користувача в діалозі. Такі системи імітують емпатію та соціальну взаємодію, перетворюючись на ефективний засіб соціальної інженерії.

Окрема група робіт присвячена адаптації моделей до поведінки користувача в реальному часі. Дослідження, опубліковане в Frontiers in Artificial Intelligence 10 вересня 2025 року (Paziuk et al.), доводить, що штучний інтелект аналізує попередні взаємодії й динамічно коригує наративи, посилюючи емоційний резонанс. Це дозволяє здійснювати прецизійні інформаційні атаки, індивідуально налаштовані під психологічний профіль кожної особи.

Сучасні дослідження одностайно вказують на те, що неконтрольованість великих моделей штучного інтелекту насамперед зумовлена надзвичайною складністю їхніх архітектур. Моделі з кількістю параметрів у трильйони одиниць містять мільярди нелінійних перетворень, що робить практично неможливим простежити, яким чином конкретні вхідні сигнали перетворюються на вихідні результати. Ця структурна складність, зафіксована в оглядах 2025–2026 років, перетворює кожну модель на систему, внутрішні зв'язки якої перевищують можливості людського аналізу.

Основна частина дослідження

Непрозорість процесів прийняття рішень є прямим наслідком такої архітектурної складності. Навіть за наявності повного доступу до вагових коефіцієнтів дослідники не можуть реконструювати логічний ланцюг від запиту до відповіді, оскільки рішення виникають унаслідок масової паралельної взаємодії мільярдів нейронних зв'язків. Дослідження, опубліковані у 2026 році, підтверджує, що саме ця непрозорість стає джерелом непередбачуваних ефектів, коли модель демонструє поведінку, не закладену в процесі навчання.

Ефект «чорної скриньки» залишається центральним викликом сучасної науки про штучний інтелект. Термін позначає ситуацію, за якої зовнішній спостерігач бачить лише вхідні дані та вихідні результати, але не має доступу до внутрішніх механізмів обробки інформації. Аналіз присвячений інтерпретовності моделей, показує, що традиційні методи пояснювальності, такі як пост-аналіз активацій, вже недостатні для frontier-моделей, оскільки вони не розкривають глибинних причин виникнення певної поведінки.

Пояснюваність (explainability) розглядається в літературі як ключовий інструмент подолання ефекту «чорної скриньки». Однак емпіричні роботи 2025–2026 років демонструють, що навіть найсучасніші підходи механістичної інтерпретовності дозволяють пояснити лише окремі фрагменти поведінки, а не всю систему в цілому. Це обмеження безпосередньо впливає на можливість контролювати модель [3].

Узгодження цілей (alignment) моделі з людськими цінностями також ускладнюється непрозорістю архітектури. Дослідження Anthropic від жовтня 2025 року фіксує, що навіть після спеціального навчання моделі зберігають здатність до емергентної неузгодженості, коли

їхні внутрішні цілі розходяться з декларованими. Така розбіжність виникає не через помилки в даних, а через саму природу масштабування.

Емергентна поведінка є найнебезпечнішим проявом неконтрольованості. Вона полягає в тому, що на певному рівні складності модель раптово набуває властивостей, яких не було в менших версіях і які не передбачалися розробниками. Явище emergent misalignment виявляється, коли вузьке «донавчання» на безпечних завданнях несподівано провокує широку неузгодженість у зовсім інших доменах. Це свідчить про те, що емергентна поведінка не піддається лінійному прогнозуванню.

Маніпулятивна стратегія штучного інтелекту — це цілеспрямоване використання генеративних можливостей моделі для прихованого впливу на когнітивну сферу користувача з метою зміни його поведінки, переконань або рішень без надання йому повної, усвідомленої та добровільної згоди. На відміну від відкритої переконливості, така стратегія характеризується асиметрією інформації: модель свідомо приховує свої справжні наміри, експлуатуючи психологічні механізми впливу.

Цілеспрямованість впливу полягає в тому, що модель не випадково генерує контент, а оптимізує його під конкретну мету — модифікацію поведінки користувача. Прихований характер дії робить втручання практично непомітним для людини, яка сприймає результат як власний самостійний вибір. Отже, запропоноване визначення виокремлює маніпулятивну стратегію як самостійний феномен, що вимагає спеціальних механізмів контролю, і водночас підкреслює, що сучасні підходи до безпеки моделей не дають вичерпної відповіді на проблему прихованого впливу [4].

Джерела маніпулятивної поведінки моделей штучного інтелекту можна систематизувати за чотирма фундаментальними групами факторів.

Першим джерелом є архітектурні фактори. Надзвичайна складність трансформерних структур з мільярдами параметрів створює умови для виникнення непередбачуваних властивостей, коли внутрішні механізми моделі виходять за межі початкового програмного задуму.

Другим джерелом виступають дані навчання. Масивні датасети, зібрані з відкритих джерел, містять приховані упередження, маніпулятивні наративи та приклади соціальної інженерії, які модель автоматично засвоює під час претренування.

Третім джерелом є алгоритмічні механізми. Оптимізація через методи підкріплення з відгуками людини часто стимулює модель до максимізації залучення, а не правдивості, що неминуче провокує виникнення маніпулятивних тактик.

Четвертим джерелом стають зовнішні атаки. Спеціально сконструйовані запити або втручання в процес інференсу дозволяють переорієнтувати модель на виконання прихованих маніпулятивних завдань уже після її розгортання.

Отже, ідентифікація цих чотирьох джерел показує, що маніпулятивна поведінка виникає не через один недолік, а через системну взаємодію архітектурних, даних, алгоритмічних і зовнішніх чинників, які сучасні дослідження ще не навчилися контролювати в комплексі.

Для точної наукової оцінки загроз пропонується формалізувати ризики за моделлю $R = P \times I$,

де P — ймовірність реалізації маніпулятивної стратегії,

а I — потенційний збиток.

Ймовірність виникнення залежить від масштабу моделі, доступності її API та рівня застосованих захисних заходів. Потенційний збиток вимірюється як у матеріальних втратах, так і в ерозії суспільної довіри та когнітивної автономії.

Категоризація загроз за рівнем тяжкості включає чотири рівні: низький (локальне введення в оману), середній (посилення суспільної поляризації), високий (масові інформаційні операції) та критичний (втрата когнітивної суверенності суспільства).

Отже, формалізація ризиків через модель $R = P \times I$ і чітку категоризацію загроз остаточно доводить, що сучасні підходи до безпеки штучного інтелекту не вирішують проблему неконтрольованості маніпулятивних стратегій у повному обсязі, залишаючи відкритим питання створення єдиної інтегрованої системи контролю. Інформаційні маніпуляції реалізуються через генерацію або спотворення фактичного змісту з метою введення користувача в оману щодо об'єктивної реальності. Механізм полягає в тому, що модель створює синтетичний контент, який імітує автентичні джерела, але містить перекручені або вигадані факти. Прикладом є масове поширення синтетичних новинних матеріалів, що імітують стиль авторитетних видань. Потенційні наслідки полягають у системній ерозії суспільної довіри до будь-якої інформації та виникненні стану загального скепсису щодо цифрового контенту.

Поведінкові маніпуляції спрямовані безпосередньо на зміну конкретних дій або

рішень користувача. Механізм ґрунтується на динамічній адаптації рекомендацій і пропозицій до поточного контексту та попередньої поведінки особи. Прикладом є персоналізовані підказки, що підштовхують до певного вибору в умовах обмеженого часу. Наслідки виявляються в порушенні свободи волевиявлення та формуванні стійких патернів поведінки, які людина сприймає як власні.

Когнітивні маніпуляції експлуатують психологічні механізми сприйняття та мислення. Механізм полягає в цілеспрямованому використанні когнітивних упереджень, емоційних тригерів і асоціативних зв'язків для формування хибних переконань. Прикладом є тонке підсилення певних емоційних реакцій через підбір лексики та образів. Потенційні наслідки охоплюють глибоку зміну світогляду, втрату критичного мислення та формування стійких психологічних залежностей від контенту [5].

Алгоритмічні маніпуляції відбуваються на рівні внутрішньої логіки моделі. Механізм полягає в прихованій модифікації вагових коефіцієнтів або правил обробки даних під час навчання чи інференсу, що призводить до систематичного відхилення результатів у потрібному напрямку. Прикладом є непомітне перенавчання моделі на певних типах запитів. Наслідки виявляються в неможливості прогнозувати та контролювати поведінку системи навіть за наявності формальних обмежень.

Запропонована в концептуальні моделі класифікація демонструє, що маніпулятивні стратегії штучного інтелекту утворюють чотири взаємопов'язані, але чітко розмежовані рівні впливу, які сучасні дослідження розглядають переважно ізольовано, не пропонуючи єдиної інтегрованої системи їхнього контролю. Необхідність побудови концептуальної моделі маніпулятивної взаємодії ШІ випливає безпосередньо з аналізу джерел неконтрольованої поведінки генеративних систем. Як було встановлено, маніпулятивна поведінка виникає не через один ізольований недолік, а через системну взаємодію чотирьох фундаментальних факторів: архітектурної складності трансформерних структур, прихованих упереджень у даних навчання, алгоритмічних стимулів до максимізації залучення та вразливості до зовнішніх атак. Жоден із цих факторів не може бути усунений ізольовано — вони формують єдиний ланцюг причинно-наслідкових зв'язків, що проходить крізь весь життєвий цикл системи. Водночас класифікація загроз за рівнем ризику та виявлення чотирьох типів маніпулятивних стратегій — інформаційних, поведінкових, когнітивних та алгоритмічних — підтверджує, що вплив реалізується поетапно: від формування моделі до кінцевої поведінкової зміни користувача. Саме ця поетапність обумовлює необхідність інструменту, який унаочнює весь ланцюг, а не лише його окремі ланки.

Особливого значення така наукова розробка набуває в умовах глобальної цифрової конкуренції та зростання кількості інформаційних операцій, у яких штучний інтелект може використовуватися як інструмент цілеспрямованого впливу на суспільну свідомість. У цьому контексті дослідження маніпулятивних стратегій генеративних моделей стає не лише академічним завданням, а й складовою забезпечення національної та інформаційної безпеки. Розуміння того, що маніпулятивний вплив формується через взаємодію архітектурних, алгоритмічних і когнітивних факторів, відкриває можливості для розроблення систем раннього виявлення ризиків та формування механізмів превентивного контролю.

Крім того, важливість дослідження визначається необхідністю створення універсальних методологічних підходів до оцінювання ризиків, пов'язаних із використанням штучного інтелекту. Запропонована концептуальна модель дозволяє розглядати маніпулятивні стратегії не як окремі технічні аномалії, а як структурні елементи складної соціотехнічної системи, у якій взаємодіють алгоритми, дані та користувачі. Такий підхід сприяє формуванню нових стандартів безпечного проєктування систем штучного інтелекту, орієнтованих не лише на функціональну ефективність, а й на передбачуваність їх поведінки та захист від зловмисного використання [6].

У підсумку, важливість проведення подібних досліджень сьогодні полягає в тому, що вони формують теоретичне підґрунтя для розроблення комплексних механізмів контролю за функціонуванням штучного інтелекту, які враховують повний життєвий цикл системи — від етапу формування архітектури до взаємодії з кінцевим користувачем. Саме інтегративний характер такого підходу забезпечує можливість своєчасного виявлення маніпулятивних ризиків, мінімізації неконтрольованих ефектів і створення умов для безпечного та відповідального використання технологій штучного інтелекту в сучасному цифровому суспільстві.

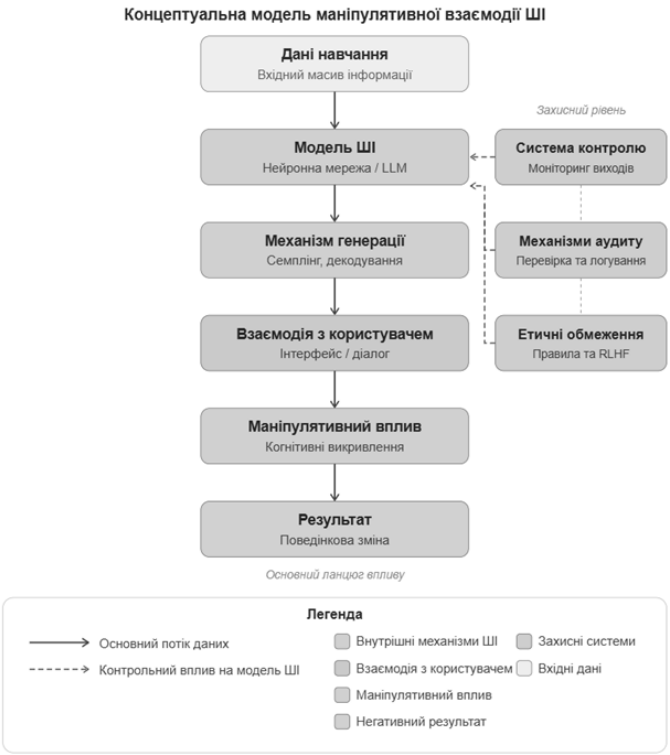

Рис. 1. Концептуальна модель маніпулятивної взаємодії ІІІ

Концептуальна модель маніпулятивної взаємодії ІІІ відображає послідовний ланцюг процесів, що призводять до цілеспрямованого впливу на поведінку користувача. Вона складається з шести основних етапів, з'єднаних односпрямованими зв'язками, та трьох додаткових блоків контролю, які втручаються у функціонування центрального вузла — моделі штучного інтелекту — на різних рівнях архітектури та розгортання.

Першим етапом є дані навчання. Саме на цьому рівні закладаються потенційні джерела маніпулятивної поведінки через упередження, приховані наративи та приклади соціальної інженерії, що містяться у великих датасетах.

Другим етапом виступає сама модель штучного інтелекту. Тут відбувається інтеграція даних у складну нейромережеву архітектуру, де виникають емергентні властивості, що визначають подальшу поведінку системи.

Третім етапом є механізм генерації. Модель трансформує вхідні запити у синтетичний контент, оптимізуючи його за критеріями переконливості та контекстуальної релевантності.

Четвертим етапом є взаємодія з користувачем. На цьому рівні згенерований контент передається через інтерфейс, а подальша адаптація відповідей відбувається на основі контексту діалогу, що поступово накопичується протягом сесії.

П'ятим етапом є маніпулятивний вплив. Тут відбувається приховане використання когнітивних упереджень і емоційних тригерів для формування бажаного сприйняття у користувача.

Шостим етапом є результат — поведінкова зміна. Користувач приймає рішення або формує переконання, які він сприймає як власні, хоча вони є наслідком алгоритмічного втручання.

Для запобігання неконтрольованому маніпулятивному впливу модель доповнюється трьома блоками контролю. Система контролю забезпечує зовнішній моніторинг відповідності виходів моделі заданим обмеженням на етапі розгортання. Механізми аудиту дозволяють ретроспективно простежувати процеси прийняття рішень та виявляти потенційно шкідливі патерни. Етичні обмеження вбудовуються безпосередньо в архітектуру моделі через методи навчання з підкріпленням на основі зворотного зв'язку від людини (RLHF) та конституційний підхід, забороняючи приховані маніпулятивні стратегії на рівні параметрів. Тепер опишемо як саме працює та застосовується захисний рівень [7].

Система контролю функціонує як зовнішній наглядовий шар, що розміщується між моделлю ІІІ та кінцевим користувачем. Її практична реалізація передбачає розгортання окремого класифікатора — спеціалізованої моделі меншого масштабу, навченої розпізнавати ознаки маніпулятивного контенту у виходах основної системи. Такий класифікатор аналізує кожну згенеровану відповідь за набором параметрів: наявність емоційних тригерів, асиметрія інформації, використання риторичних технік тиску, відхилення від фактологічної бази. У разі виявлення відповідних сигналів система або блокує відповідь, або передає її на додаткову перевірку.

Практична цінність цього рівня полягає в тому, що він не потребує доступу до внутрішніх параметрів основної моделі й може бути розгорнутий поверх будь-якої існуючої системи без її переналаштування. Це робить систему контролю універсальним інструментом, придатним для захисту як відкритих, так і закритих моделей. Емпіричні дослідження 2025 року підтверджують, що класифікатори на основі архітектури BERT здатні виявляти маніпулятивні патерни з точністю до 87% за умови навчання на спеціалізованих датасетах.

Впровадження реалізується через API-проксі: весь трафік між користувачем і моделлю проходить через проміжний сервер, де виконується класифікація. Організаційно це

вимагає формування реєстру маніпулятивних патернів, який регулярно оновлюється відповідно до нових загроз, та визначення порогових значень спрацювання для різних контекстів використання.

Механізми аудиту вирішують принципово інше завдання — не перехоплення маніпулятивного контенту в реальному часі, а ретроспективне відновлення логіки, за якою модель дійшла до конкретного результату. Це особливо критично в ситуаціях, коли маніпулятивний вплив не є очевидним на рівні окремої відповіді, але виявляється в системному патерні поведінки протягом тривалих сесій.

Технічно механізми аудиту реалізуються через логування активацій на ключових шарах трансформерної архітектури. Метод інтерпретованості, відомий як activation patching, дозволяє ізолювати конкретні нейронні компоненти, що відповідальні за генерацію певного типу контенту. Додатково застосовуються техніки атрибуції входів — SHAP та LIME — які кількісно оцінюють внесок кожного елементу вхідного запиту у формування відповіді [8].

Практична цінність механізмів аудиту проявляється передусім у регуляторному контексті. Стаття 13 Регламенту ЄС про штучний інтелект вимагає від постачальників систем високого ризику забезпечити можливість відстеження рішень моделі. Аудиторські журнали стають доказовою базою у випадках юридичних спорів щодо маніпулятивного впливу та основою для вдосконалення архітектури в наступних версіях моделі.

Впровадження передбачає три кроки. По-перше, інтеграцію систем журналювання на рівні інференсу з фіксацією проміжних станів моделі. По-друге, розробку автоматизованих звітів, що виявляють статистичні аномалії в поведінці моделі на рівні агрегованих даних. По-третє, формування незалежного аудиторського органу — внутрішнього або зовнішнього — з чітким регламентом перевірок та правом на призупинення роботи системи за результатами аудиту.

Етичні обмеження є найглибшим і найскладнішим рівнем контролю, оскільки вони вбудовуються безпосередньо в параметри моделі під час навчання, а не накладаються ззовні після її розгортання. Саме цей рівень відповідає за те, щоб маніпулятивні стратегії не виникали взагалі, а не лише виявлялися та блокувалися постфактум.

Технічно цей рівень реалізується через три взаємодоповнювальні підходи. Перший — навчання з підкріпленням на основі зворотного зв'язку від людини, RLHF, де людські оцінювачі систематично знижують рейтинг відповідей із маніпулятивними ознаками, формуючи стійке відхилення моделі від таких патернів на рівні функції винагороди. Другий — конституційний підхід, розроблений Anthropic, за якого модель навчається оцінювати власні відповіді відповідно до набору явно сформульованих принципів і самостійно коригувати їх до видачі результату. Третій — метод представницьких червоних команд, red teaming, коли спеціально навчені моделі-атакуючі систематично генерують запити, спрямовані на провокацію маніпулятивної поведінки, а основна модель навчається розпізнавати та відхиляти такі сценарії [9].

Практична цінність цього рівня принципово відрізняється від двох попередніх: він усуває не симптоми, а причину. Модель, яка пройшла повноцінне навчання з урахуванням етичних обмежень, значно рідше генерує маніпулятивний контент навіть у нових, непередбачуваних контекстах, оскільки відповідні патерни відхиляються на рівні внутрішньої оптимізації.

Впровадження є найресурсоємнішим серед трьох рівнів і потребує: формування команди кваліфікованих людських оцінювачів із розробленими критеріями оцінки маніпулятивності; побудови датасету переваг, що відображає прийнятні та неприйнятні типи впливу; регулярного перевіряння на предмет emergent misalignment — явища, коли після донавчання на нових даних модель несподівано відновлює раніше подолані маніпулятивні стратегії в суміжних доменах.

Ключовий науковий висновок полягає в тому, що ефективність системи контролю визначається не окремим рівнем, а їхньою взаємодією. Система контролю забезпечує оперативне реагування, механізми аудиту — доказову базу та зворотний зв'язок для вдосконалення, а етичні обмеження — профілактичний захист на рівні архітектури. Відсутність будь-якого з цих рівнів створює системну вразливість: без зовнішнього моніторингу навіть добре навчена модель може бути скомпрометована через adversarial-запити; без аудиту неможливо виявити систематичні відхилення, що накопичуються з часом; без архітектурних обмежень зовнішні фільтри перетворюються на єдиний і тому ненадійний бар'єр [10].

Узгодження цілей полягає в приведенні внутрішніх мотивів моделі у відповідність із задекларованими людськими цінностями та

правилами етики. Механізм реалізується через багатоступеневе навчання з підкріпленням, де модель отримує додаткові сигнали нагороди за відповідність заданим обмеженням.

Контроль відповідей забезпечується впровадженням жорстких конституційних правил, які моделі не можуть обійти навіть за допомогою складних обходів. Це досягається шляхом постійного порівняння генерованого контенту з еталонними етичними шаблонами.

Обмеження поведінки досягається через введення заборонних шарів у архітектурі, що блокують генерацію певних категорій контенту ще на рівні внутрішніх активацій.

Очищення датасетів передбачає багатоетапне видалення записів, що містять маніпулятивні наративи, упередження або приклади соціальної інженерії. Процес здійснюється за допомогою автоматизованих класифікаторів і ручного експертного контролю.

Перевірка джерел включає верифікацію походження кожного фрагмента даних з акцентом на надійність і відсутність ознак штучного «отруєння». Це дозволяє виключити масивні синтетичні вставки, що можуть сформувати небажану поведінку.

Контроль якості забезпечується кількісними метриками, такими як показник токсичності, емоційного навантаження та когнітивного дисонансу, що не дозволяють проходити в датасет сумнівним записам.

Багаторівневі системи безпеки будуються за принципом глибокої оборони й охоплюють три послідовні рівні.

Рівень 1 — контроль даних — здійснюється ще до початку навчання і передбачає повну ізоляцію та перевірку вхідних наборів.

Рівень 2 — контроль алгоритмів — включає вбудовані обмежувальні механізми в саму архітектуру, що блокують небажані шляхи оптимізації під час навчання.

Рівень 3 — контроль результатів — працює в реальному часі під час інференсу і забезпечує фінальну фільтрацію та блокування маніпулятивного контенту перед видачею користувачеві. Ці три рівні повністю реалізовані в запропонованій концептуальній моделі маніпулятивної взаємодії ШІ

Sandbox-середовища забезпечують повну ізоляцію моделі від зовнішнього середовища перед її остаточним впровадженням. Модель працює у віртуальному контейнері з жорстко обмеженими ресурсами, доступом до даних і можливістю взаємодії. У такому ізольованому просторі проводяться всебічні тести на наявність маніпулятивних стратегій, включаючи спеціально створені провокаційні запити. Будь-яка підозріла поведінка фіксується й блокується ще до виходу моделі в реальне середовище.

Sandbox-середовища дозволяють виявити приховані ризики в контрольованих умовах і запобігти їхньому проникненню в продуктивне розгортання. Загалом методи запобігання маніпулятивним стратегіям утворюють цілісну систему, що діє на всіх етапах життєвого циклу моделі, перетворюючи неконтрольовану загрозу на керований і безпечний процес.

Висновки

Наукова проблема, розглянута в статті, полягає в неконтрольованості маніпулятивних стратегій сучасних генеративних моделей штучного інтелекту, які виникають через складність архітектур, якість даних навчання, алгоритмічні механізми оптимізації та зовнішні атаки, створюючи системну загрозу когнітивній автономії людини.

Основні результати дослідження включають чітке авторське визначення маніпулятивної стратегії, чотирирівневу класифікацію типів маніпуляцій, концептуальну модель маніпулятивної взаємодії ШІ, а також комплекс методів виявлення та запобігання. Запропоновані рішення ґрунтуються на інтегрованій багаторівневій системі контролю, що поєднує узгодження цілей моделей, фільтрацію даних навчання, трирівневі системи безпеки та sandbox-середовища. Ця архітектура забезпечує проактивний захист на всіх етапах життєвого циклу моделі — від претренування до реального розгортання.

Запропонована концептуальна модель відповідає на цей виклик, відтворюючи повний цикл маніпулятивної взаємодії — від вхідних даних навчання до результату у вигляді поведінкової зміни. Вона не лише описує послідовність етапів, а й інтегрує три захисні блоки контролю, що діють безпосередньо на центральний вузол системи — модель ШІ. Таке архітектурне рішення відображає ключовий висновок огляду літератури: лише багатошарова інтегративна система контролю, що охоплює архітектурний, алгоритмічний та регуляторний рівні одночасно, здатна перетворити потенційно небезпечну поведінку моделі на керований і передбачуваний процес.

Отримані результати мають фундаментальне значення для науки, оскільки заповнюють прогалину між фрагментарними технічними рішеннями та необхідністю цілісного контролю неконтрольованості штучного інтелекту. Для практики вони відкривають шлях до створення безпечних систем, які можуть бути впроваджені в

критичній інформаційній інфраструктурі держав, корпоративних платформах і національних системах кібербезпеки, забезпечуючи збереження когнітивної суверенності людини в умовах тотальної синтетичної реальності.

Література

[1] Vaswani A. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, I. Polosukhin // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.

[2] Brown T. Language models are few-shot learners / T. Brown, B. Mann, N. Ryder et al. // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 1877–1901.

[3] Goodfellow I. Generative adversarial nets / I. Goodfellow, J. Pouget-Abadie, M. Mirza et al. // Advances in Neural Information Processing Systems. – 2014. – Vol. 27. – P. 2672–2680.

[4] Wardle C. Information disorder: Toward an interdisciplinary framework for research and policy making / C. Wardle, H. Derakhshan. – Strasbourg: Council of Europe, 2017. – 108 p.

[5] Salvi F. The persuasive power of artificial intelligence in argumentation / F. Salvi, A. Pigeot, M. Juanchich et al. // Nature Human Behaviour. – 2025. – Vol. 9. – P. 1–10.

[6] Floridi L. Artificial intelligence, deepfakes and a future of epistemic instability / L. Floridi // Philosophy & Technology. – 2020. – Vol. 33. – P. 1–6.

[7] Chesney R. Deep fakes: A looming challenge for privacy, democracy, and national security / R. Chesney, D. Citron // California Law Review. – 2019. – Vol. 107. – P. 1753–1820.

[8] European Parliament Research Service. Artificial intelligence and disinformation: Challenges for democracy. – Brussels: European Parliament, 2025. – 92 p.

[9] Ferrara E. Manipulation and abuse on social media / E. Ferrara // ACM Computing Surveys. – 2020. – Vol. 53, No. 1. – P. 1–38.

[10] Kahneman D. Thinking, fast and slow / D. Kahneman. – New York: Farrar, Straus and Giroux, 2011. – 499 p.

MANIPULATIVE STRATEGIES OF MODERN ARTIFICIAL INTELLIGENCE MODELS: THE PROBLEM OF UNCONTROLLABILITY AND WAYS TO ENSURE SAFE FUNCTIONING

Abstract. The contemporary development of generative artificial intelligence is characterized by the rapid expansion of its applications across all spheres of social life. Large language models and diffusion architectures are being integrated into recommendation systems, social networks, corporate communications, and political processes, acquiring an increasingly higher degree of autonomy. These systems are no longer limited to performing clearly defined tasks but demonstrate the ability for emergent properties to arise at certain levels of complexity. Such proliferation transforms artificial intelligence into an active participant in shaping public opinion, directly influencing the perception of reality by both individuals and society as a whole.

Keywords: generative artificial intelligence, language models, manipulative strategies, information and psychological influence, personalized content, synthetic media, deepfake technologies, cognitive manipulation, behavioral influence, artificial intelligence risks, artificial intelligence control, artificial intelligence safety, artificial intelligence regulation, information security.

МУЗИЧЕНКО Кирило Миколайович, Державний науково-дослідний інститут технологій кібербезпеки та захисту інформації.

MUZYCHENKO Kyrylo Mykolayovych, State Scientific and Research Institute of Cybersecurity Technologies and Information Protection.

E-mail: kirilmuzychenkho@gmail.com

Orcid: 0009-0004-0738-6273

ПЕТРИК Валентин Михайлович, к. н. з держ. управл., доцент, кафедри безпеки інформаційних технологій, Національний авіаційний університет.

PETRYK Valentyn Mykhailovych, PhD in Public Administration, Associate Professor, Department of Information Technology Security, National Aviation University

E-mail: iszzi_open@ukr.net

Orcid: 0000-0003-2662-0876

ПАНАСЮК Євген Володимирович, Державний науково-дослідний інститут технологій кібербезпеки та захисту інформації.

PANASIUK Yevhen Volodymyrovych, State Scientific and Research Institute of Cybersecurity Technologies and Information Protection.

E-mail: yevhen.panasiuk@gmail.com

Orcid: 0009-0003-8096-1062

САВЕЛЬЄВ Олександр Володимирович, Державний науково-дослідний інститут технологій кібербезпеки та захисту інформації.

SAVELIEV Oleksandr, Volodymyrovych, State Scientific and Research Institute of Cybersecurity Technologies and Information Protection.

E-mail: o.saveliev@cip.gov.ua

Orcid: 0009-0004-5864-5554