DOI: https://doi.org/10.36059/978-966-397-661-7-20

АСПЕКТИ ПОБУДОВИ НОВИХ СТРАТЕГІЙ ЗАХИСТУ ЦИФРОВИХ СИСТЕМ ПІДТРИМКИ ОСВІТНЬОГО ПРОЦЕСУ У МІНЛИВОМУ ЛАНШАФТІ «КОГНИТИВНИХ ВІЙН»

Макаров Олексій Вікторович,
старший викладач кафедри хімічних технологій
Навчально-наукового інституту
хімічних технологій та фармації
Національного університету «Одеська політехніка»,
пр. Шевченка, 1, м. Одеса, 65044, Україна

Вступ

Наш час характеризується безпрецедентною конвергенцією глибокої цифрової трансформації та еволюцією парадигм нелінійних конфліктів, що найвиразніше втілено в новітній концептуальній моделі «когнітивної війни». Оскільки системи вищої освіти та професійного навчання все більше покладаються на складні, взаємопов'язані цифрові архітектури – від хмарних систем управління навчанням (Cloud LMS) та педагогічних асистентів на базі штучного інтелекту (artificial intelligence, AI) до розподілених репозиторіїв знань – вони мимоволі розширюють площину своїх вразливостей. У цьому мінливому ландшафті цифрові системи підтримки освітнього процесу перестають бути просто адміністративними чи логістичними інструментами; вони трансформувалися в критичні вектори національної та інституційної когнітивної стійкості. Відповідно, захист цих систем вимагає фундаментального відходу від традиційних, реактивних моделей кібербезпеки в бік динамічних, предиктивних та багаторівневих стратегій захисту.

Специфіка та висока чутливість сектору професійного навчання суттєво загострюють нагальність цієї зміни парадигми. На відміну від загальної освіти, середовища професійної підготовки несуть безпосередню відповідальність за формування ключових компетентностей персоналу критичної інфраструктури, інженерів, державних службовців, військових фахівців та осіб, які приймають рішення в корпоративному секторі. Порушення цілісності, доступності або автентичності

даних у цих спеціалізованих освітніх каналах може призвести до катастрофічних каскадних наслідків на наступних етапах. Зловмисне втручання або несанкціонована модифікація професійних навчальних програм виявляються не просто технічним простоєм; вони безпосередньо загрожують епістемологічній безпеці держави. Шляхом потенційної інституціоналізації хибних методологій, підриву довіри до сертифікованих знань або прихованого маніпулювання матрицями прийняття рішень майбутніх експертів, такі атаки спрямовані на саму основу професійної компетентності. Саме це перетинання вразливості цифрової інфраструктури та когнітивного маніпулювання становить суть сучасних векторів асиметричних загроз.

Крім того, ескалація когнітивної війни докорінно змінює саму природу загрози. Кіберпротивники більше не обмежуються традиційними технічними експлойтами, такими як розподілені атаки на відмову в обслуговуванні (DDoS), програми-вимагачі або несанкціоноване вилучення даних. Замість цього ворожі актори все частіше використовують складні, комбіновані операції, де технічні кіберзлами слугують інфраструктурним інструментом для когнітивної деградації. Маніпулюючи освітнім контентом, змінюючи метрики оцінювання або впроваджуючи синтетичні дані в цифрові навчальні середовища, противники можуть здійснювати кампанії підсвідомого впливу, провокувати інституційну дестабілізацію та знижувати колективний інтелектуальний потенціал цільових когорт.

У такому нестабільному середовищі несвоєчасне реагування на інциденти безпеки є еквівалентом системного збою. Затримка у виявленні та усуненні загроз дозволяє зловмисним ін'єкціям безперешкодно поширюватися освітньою екосистемою, тим самим нормалізуючи спотворену інформацію та підриваючи фундаментальну довіру, необхідну для ефективного педагогічного процесу.

Тому створення своєчасних, адаптивних та стійких механізмів реагування є першорядною умовою для забезпечення безперервності та достовірності професійного навчання. Нові оборонні стратегії повинні інтегрувати виявлення аномалій у режимі реального часу, криптографічну валідацію освітнього контенту та когнітивну телеметрію для фіксації ледве помітних поведінкових або інформаційних відхилень. Отже безпека цифрових систем підтримки освітнього процесу нерозривно пов'язана із збереженням когнітивного суверенітету та стратегічної інституційної стабільності в епоху перманентного гібридного протистояння.

1 Огляд зсуву кіберзагроз у сучасному тренді використання AI

Стрімкий розвиток технологій штучного інтелекту (AI), зокрема великих мовних моделей (LLM) та генеративних змагальних мереж, призвело до фундаментального зрушення у парадигмі цифрової взаємодії людини та цифрових систем протягом дуже короткого часу. Якщо раніше соціальна інженерія та методи онлайн-маніпуляції спиралися на статичні, відносно прості у розпізнавані шаблони та масові розсилки, то сьогодні технології AI сприяють виникненню нового рівня небезпеки – «індустріалізованого обману» (industrialized deception). Інтеграція принципів когнітивної психології та машинного навчання відкрила безпрецедентні можливості для прихованого впливу, дозволяючи генерувати переконливий, гіперперсоналізований синтетичний контент у великих масштабах та з високою швидкістю. Саме в цьому контексті з'явився термін «когнітивна війна» – вид збройного чи гібридного конфлікту, в якому полем бою є людський розум[765]. Мета такої війни – змінити те, як люди думають та діють, підірвати довіру до адміністративних інституцій, придушити волю до опору та трансформувати базові цінності супротивника. Сучасні мовні моделі здатні автоматично аналізувати цифровий слід користувача виявлення його психологічних характеристик, диспозицій і нахилів. Це значно знизило бар'єр входу щодо витончених системних гібридних атак. Зловмисники використовують LLM для створення масштабних цілеспрямованих кампаній, які майстерно експлуатують когнітивне сприйняття людини. AI-згенеровані тексти, що апелюють до помилкового авторитету або створюють почуття надзвичайної терміновості, виявляються дуже успішними в шахрайських схемах при обході критичного мислення та когнітивного сприйняття жертви.

У цьому контексті традиційні засоби кіберзахисту, такі як спам-фільтри або системи виявлення вторгнень, що працюють на основі сигнатур, демонструють повну неефективність. Синтаксично ідеальний і контекстуально релевантний текст, згенерований AI, вимагає переходу до нового підходу – виявлення аномалій (behavioral anomaly detection), оскільки алгоритмічні маніпуляції все частіше маскуються під звичні людські психологічні тактики і стилі комунікації. Цифрова психологія обману не обмежується лише текстовими комунікаціями; вона глибоко інтегрується в цифрові інтерфейси (UX). Використання AI призвело до появи

[765] Oriyano S. Deceptive Intelligence: AI, Social Engineering, and Securing the Human Element. New York : Apress, 2026. 727 p. C. 4. ISBN 979-8-8688-2176-9

нових адаптивних «темних патернів», які швидко підлаштовуються під поведінку конкретного користувача в режимі реального часу[766].

На відміну від статичних хитрощів (наприклад, заздалегідь зазначених галочок або прихованих платежів), AI-керовані інтерфейси стали в потрібний момент застосовувати нові тактики штучного дефіциту (Artificial Scarcity) або «пастки» (Roach Motel), навмисно ускладнюючи процес скасування рішення або відмови від послуги саме тоді, коли система фіксує. Ці елементи дизайну використовують фундаментальні психологічні тенденції, щоб непомітно підштовхнути людину до рішень, які вона на планувала приймати.

На макрорівні генеративний AI породив феномен, який отримав назву «слопаганда» – масове виробництво та поширення низькоякісного, але емоційно залучающого контенту, спрямованого на маніпуляцію суспільною свідомістю[767]. Слопатанда поєднує в собі нульову вартість алгоритмічної генерації «інформаційного сміття» з класичними техніками мікротаргетованої пропаганди.

Успіх AI-маніпуляцій багато в чому обумовлений складною психологічною динамікою довіри між людиною та машинним інтерфейсом. Дослідження виявляють парадокс генеративного AI: хоча люди часто не здатні відрізнити синтетичний текст або зображення від реального, лише розкриття інформації про авторство AI може різко знизити рівень довіри до правдивого контенту[768]. Більше того, люди схильні некоректно оцінювати судження AI у стресових ситуаціях. Експерименти показують, що користувачі часто помилково відкидають об'єктивно правильні висновки машини, якщо алгоритм діє занадто впевнено або висуває обвинувальні судження (accusatory judgments). З іншого боку, AI-системи, спеціально навчені використовувати елементи соціальної інженерії та симулювати емпатію для пояснення своїх дій, часто виявляються переконливішими, ніж чесні системи, що ще більше вкорінює віру в дезінформацію[769]. Ця прірва між

[766] Pandey D. Emergent Dark Patterns in AI-Generated User Interfaces : arXiv preprint arXiv:2602.18445. Mathura : GLA University, 2026. 15 p. C. 1, 8. DOI: https://doi.org/10.48550/arXiv.2602.18445

[767] Klincewicz M., Alfano M., Fard A. E. Slopaganda: The interaction between propaganda and generative AI. Filosofiska Notiser. 2025. Årg. 12, Nr. 1. P. 135–162. C. 138. DOI: https://doi.org/10.48550/arXiv.2503.01560

[768] Loth A., Kappes M., Pahl M.-O. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)*, April 13–17, 2026, Dubai, United Arab Emirates. New York : ACM, 2026. P. 1–8. C. 2, 4. DOI: https://doi.org/10.1145/3774905.3795471

[769] Loconte R., Monaro M., Pietrini P., Verschuere B., Kleinberg B. Humans incorrectly reject confident accusatory AI judgments. *Computers in Human Behavior*. 2026. Vol. 182. C. 15. DOI: https://doi.org/10.1016/j.chb.2026.108530

суб'єктивним сприйняттям та фактичною точністю (perception-accuracy gap) оголює нашу критичну когнітивну вразливість перед цифровими агентами.

Впровадження LLM та синтетичних медіа (таких як аудіо- та відео-дипфейки) у академічний, політичний, соціальний дискусії є прямою загрозою для розвитку демократичних принципів та інституційної довіри. Особливу небезпеку таять у собі «мікрофейки» (microfakes) – гіперлокалізований синтетичний контент, який оминає традиційні системи перевірки фактів, сіє плутанину на локальному рівні та посилює соціальну поляризацію[770].

У такій ситуації протидія цим загрозам вимагає усвідомлення того факту, що суто технічні методи виявлення приречені на відставання через постійну гонку озброєнь (arms race) між генеративними моделями і детекторами. Ефективне рішення потребує застосування комплексного системного підходу, що розглядає дипфейки та AI-алгоритми як складні соціо-технічні конструкти, вбудовані у повсякденне життя суспільства.

Для забезпечення інформаційної безпеки необхідна розробка єдиної нормативно-правової бази (такої як адаптація положень AI Act), впровадження криптографічних стандартів підтвердження походження цифрового контенту (наприклад, C2PA), а також повсюдне підвищення рівня медійної та алгоритмічної грамотності (AI literacy) громадян[771]. Тільки завдяки об'єднанню зусиль у галузі поведінкового аналізу, етичного регулювання та захисту фундаментальних прав суспільство зможе вибудувати ефективний захист проти витонченої алгоритмічної індустрії цифрового обману.

2 Історія розвитку цифрової підтримки навчального процесу та фіксації інцидентів на прикладі Інституті хімічних технологій та фармації

Стійкий тренд на цифрову трансформацію навчального процесу особливо в прикладних технічних курсах став проявлятися з масовим поширенням персональних ЕОМ та мережі Інтернет. Першою цифровою сервісною функцією для здобувачів вищої освіти стало розміщення на сайті інституту конспектів лекцій та матеріалів для лабораторних та практичних робіт у вільному доступі для ознайомлення та скачування.

[770] Yousefi Y., McIntyre A., Russo F., Sanchez Galera M. D., Tonello T. Regulatory innovations and policy options for generative AI and digital democracy GANs : Deliverable D6.2. *SOLARIS Project (HORIZON Grant Agreement 101094665)*. 2026. 48 p. С. 18. URL: https://projects.illc.uva.nl/solaris

[771] Там само. С. 24, 38.

Інформація розміщувалася як фали архівів з доступом до скачування з веб-сторінок з описом вмісту або повністю стилізовані для самостійної роботи в мережі Інтернеті методичні матеріали на web-сервері інституту. Незважаючи на ранню епоху розвитку мережевих технологій співробітниками технічної підтримки університетських цифрових ресурсів вже стали фіксуватися інциденти різного рівня складності та наслідків. Такі інциденти поступово, але послідовно трансформували відгуки на них відповідно до глибини вразливостей та економічної ефективності заходів у відповідь: обов'язкова авторизація та журналування при мережному друку після «жарту» з мережевим принтером; відмова від корпоративного sendmail-сервісу на платформі Novell NetWare на користь онлайн поштових служб з web-інтерфейсом після тривалої боротьби зі зломами та спам-розсилками; обов'язкова дво-факторна авторизація при доступі користувача до інститутського облікового запису з нового пристрою після проблем з викраденням облікових записів та особистої інформації користувачів та таке інше. Згодом такі фактори як пандемія, онлайн викладання, аварійні відключення світла, пряме руйнування корпусу внаслідок попадання дрону призвели до переходу на хмарні послуги у всіх основних стратегіях цифрової підтримки освітнього процесу. Зовнішні провайдери гарантують безперебійну роботу і цілодобовий доступ до розміщеного контенту і цифрових продуктів. Однак інститут підтримує власну мережеву архітектуру: периметр науково-дослідних серверів на базі Ubuntu та інститутський репозиторій бібліотечних та науково-дослідних сервісів для постійного оперативного застосування у поточних навчально-методичних та науково-дослідних процесах, комерційних впроваджувальних проектах та підвищенні кваліфікації. Така гібридна зв'язка локальних ресурсів із хмарними покриває всі необхідні стратегії ведення навчального процесу з науково-практичною діяльністю, є ефективною в епонімічному плані, висвітлює основні виклики кібербезпеки. Слід зазначити, що співробітники інституту хімічної технології та фармації Національного університету «Одеська політехніка» завжди з великим ентузіазмом відстежували та впроваджували цифрові інновації завдяки участі у дослідницьких спільнотах та групах різних цифрових компаній. Серед них Microsoft, Cisco, Autodesk, Oracle, Canonical та багато інших. Паралельно з технологіями реалізації цифрових можливостей удосконалення навчального процесу в інституті завжди активно досліджувалися сфери ефективності цифрового викладання та рівня залучення слухачів до курсів лекцій із застосуванням тих чи інших стратегій ведення освітнього процесу. Глибокий аналіз міжнародного досвіду

впровадження платформ цифрового викладання та підходів до вимірів освітнього результату таких впроваджень дозволяв послідовно тестувати найпередовіші методології зняття метрик навчальної діяльності здобувачів у реальному часі проведення навчального процесу, аналізувати цифрові результати з навичками вирішення практичних та екзаменаційних робіт для розгляду професійних траєкторій випускників та актуалізації навчального процесу.

3 Перехід від «Системи управління» до «Системи партнерства»

Результатом переліченої вище діяльності виявилася основна ідея підтримки високого рівня залученості претендентів з одночасним збільшенням навчального навантаження з основних дисциплін: цифрова система цифрової підтримки освітнього процесу має базуватися на принципі студентів як партнерів у освітньому процесі інституту. Замість жорсткої функції контролю проходження здобувачем запланованих навчальною програмою курсових активностей, система має реалізовувати ініціативи щодо мотивації та саморегуляції здобувача в освітньому циклі. Як показала безперервна практика проведення вимірювань показників академічної успішності та задоволеності отриманою освітньою послугою ключовим аспектом стала інтеграція в освітній процес стратегій командної роботи, гнучких методологій та інструментів для співтворчості (co-creation) у реалізації навчальних курсів, зовнішніх змагальних заходів і додаткових активностей (кружки, семінари, консультації).

Саме вибудовування довірчих командних комунікацій з перших днів занять та протягом усього часу навчання, глибоке розуміння гнучких методологій командної роботи на величезному масиві реальних прикладів з освітніх ініціатив та зовнішніх конкурсних завдань, застосування та практичне засвоєння найсучасніших цифрових інструментів ведення професійної діяльності дали позитивні результати освітньої діяльності, особливий рівень відповідальності за використання корпоративних ресурсів, високі оцінки ведення курсів та високу задоволеність освітнім середовищем. Заплановані технології інновацій найближчим часом пов’язані в інституті з максимальною інтеграцією в цифрові освітні платформи інтелектуальних агентів підтримки комунікації у реальному режимі (Chatbots & AI Assistants). Платформи включатимуть адаптивних чат-ботів, здатних підтримувати довгостроковий контекст спілкування

(multi-session state modeling)[772]. Такі технології дозволяють платформі доповнити якісне наставництво (mentorship) викладача та розвантажити трафік його комунікацій за курсовими та проектними завданнями завдяки збереженню, обробці та ефективному застосуванню довгострокової «пам'яті» взаємодії.

4 Безпека та верифікація в умовах AI-децепції: комплексний аналіз механізмів, загроз та стратегій захисту

Стрімкий розвиток великих мовних моделей (LLM) і мультимодальних генеративних систем ознаменував фундаментальне зрушення в ландшафті кібербезпеки, перетворивши соціальну інженерію з ручного процесу, що відносно легко виявляється в масштабовану індустрію обману (industrialized deception)[773]. Якщо раніше безпека будувалася на виявленні технічних аномалій (сигнатур шкідливого програмного забезпечення, фішингових посилань, синтаксичних помилок), то сьогодні атаки змістилися на когнітивний рівень. Як зазначає Феррара у концепції «Парадоксу генеративного AI» (Generative AI Paradox), у міру того як синтетичний контент стає візуально та семантично невідмінним від справжнього, соціум починає раціонально знецінювати будь-які цифрові докази, що веде до системної ерозії довіри до інформаційної інфраструктури. У цих умовах традиційні методи верифікації, що ґрунтуються на людському сприйнятті, втрачають свою ефективність, вимагаючи переходу до криптографічних, поведінкових та інституційних механізмів захисту[774].

Розглянемо архітектуру безпеки та верифікації в умовах AI-децепції (AI-децепція, обман штучним інтелектом – це здатність AI-систем навмисно вводити користувачів, розробників або перевіряючі алгоритми в оману), спираючись на новітні академічні дані, що охоплюють психологічні вектори атак, технічні алгоритми виявлення аномалій, проблеми гібридного (людино-машинного) прийняття рішень та регуляторні інновації.

В умовах AI-децепции головним вектором атаки стає не програмний код, а людська психіка. Інтеграція методів когнітивної психології

[772] Oriyano S. Deceptive Intelligence: AI, Social Engineering, and Securing the Human Element. New York : Apress, 2026. 727 p. C. 341, 342. ISBN 979-8-8688-2176-9

[773] Loth A., Kappes M., Pahl M.-O. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)*, April 13–17, 2026, Dubai, United Arab Emirates. New York : ACM, 2026. P. 1–8. C. 1. DOI: https://doi.org/10.1145/3774905.3795471

[774] Там само. P. 1–8. C. 2.

та машинного навчання дозволяє зловмисникам автоматизувати збір відкритих даних (OSINT) для створення «гіперперсоналізованих» профілів жертв. Дослідження показують, що сучасні LLM здатні з високою точністю аналізувати цифровий слід користувача та робити висновки про його психологічні диспозиції, ґрунтуючись на моделі «Великої п'ятірки» (Big Five: відкритість, сумлінність, екстраверсія, доброзичливість, нейротизм)[775]. На основі цих даних алгоритми генерують фішинговий контент, що ідеально резонує з емоційним станом жертви. У ході симуляцій було доведено, що атаки фішингу, адаптовані під особистісні особливості жертви за допомогою LLM, досягають теоретичного показника клікабельності (click-through rate), який на 40 % перевищує показники стандартного фішингу.

Зловмисники експлуатують так звані «темні патерни» (Dark Patterns) – елементи інтерфейсу та комунікації, які маніпулюють користувачем, спираючись на штучний дефіцит, хибний авторитет чи соціальний доказ[776]. Більш того, AI впроваджує механізми «багатосеансового моделювання стану» (multi-session state modeling). Персистентні чат-боти зберігають безперервність контексту в розірваних у часі розмовах, використовуючи векторні бази даних (такі як Weaviate чи Pinecone)[777]. Бот запам'ятовує деталі, згадані жертвою, створюючи безшовну ілюзію емпатії та людської пам'яті. Це дозволяє обійти критичне мислення користувача, оскільки довіра формується не через агресивний тиск, а через рутинну, безпечну взаємодію. Такий патерн називається вектори психологічної поверхні. (Psychological Surface Vectors).

Фундаментальна проблема верифікації тепер полягає в тому, що сприйняття людини більше не є абсолютно надійним сенсором. Емпіричні дані свідчать про існування когнітивного «розриву між сприйняттям та фактичною точністю»[778]. Експерименти показують, що навіть коли користувачі виявляють підвищену зосередженість та підозрілість, їхня здатність

[775] Rostova E. V. Psychological Surface Vectors: Mitigating Large Language Model-Driven Social Engineering via Behavioral Anomaly Detection. *The American Journal of Engineering and Technology*. 2025. Vol. 7, № 10. P. 178–184. C. 178. URL: https://www.theamericanjournals.com/index.php/tajet/article/view/6944

[776] Pandey D. Emergent Dark Patterns in AI-Generated User Interfaces : arXiv preprint arXiv:2602.18445. Mathura : GLA University, 2026. 15 p. C 1, 8. DOI: https://doi.org/10.48550/arXiv.2602.18445

[777] Oriyano S. Deceptive Intelligence: AI, Social Engineering, and Securing the Human Element. New York : Apress, 2026. 727 p. C. 341, 342. ISBN 979-8-8688-2176-9

[778] Loth A., Kappes M., Pahl M.-O. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26), April 13–17, 2026, Dubai, United Arab Emirates. New York : ACM, 2026. P. 1–8. C. 5. DOI: https://doi.org/10.1145/3774905.3795471

визначати згенерований AI-контент залишається на рівні випадкового вгадування. Більше того, асиметрична когнітивна втома (cognitive fatigue), що виникає при тривалому впливі синтетичного контенту, знижує точність виявлення фейків людиною до 10,20 %. Назва таких векторів атак називають як феномен «розриву між сприйняттям і точністю» (Perception-Accuracy Gap).

Спроби вирішити проблему AI-децепції виключно технічними методами (розробка детекторів дипфейків або аналізаторів тексту) призводять до класичної «гонки озброєнь» між генеративними та детектируючими моделями (Generative-Detection Arms Race)[779].

Традиційні системи виявлення вторгнень та спам-фільтри, що працюють на основі пошуку ключових слів або синтаксичних сигнатур, виявляються безсилими перед синтаксично ідеальним та контекстуально релевантним текстом LLM. Подібні системи здатні заблокувати лише близько 34 % таких атак[780]. Зловмисники застосовують техніки приховування намірів (jailbreaking), такі як «злом ланцюжка міркувань» (chain-of-thought injection) або впровадження шкідливих цілей у доброякісні наративи.

Детектори дипфейків, що аналізують артефакти стиснення, розсинхронізацію губ (lip-sync drift) або спектральні аномалії голосу, також втрачають ефективність. Сучасні генеративні конвеєри оптимізують приховані ознаки (latent features) спеціально для обходу детекторів. Як зазначає Оріяно, чистий, згенерований з частотою 30 кадрів в секунду, дипфейк, переданий через протокол з агресивним стисненням відео (наприклад, Zoom або Microsoft Teams), згладжує піксельні артефакти, обходячи більшість систем виявлення з відкритим вихідним кодом. «Візуальна точність стає нерелевантною, якщо всі зовнішні системи координат збігаються з тим, що користувач вважає реальним»[781].

Оскільки верифікація контенту на рівні артефактів більше не дає абсолютних гарантій, безпека автоматично зміщується у бік аналізу поведінки та намірів (Behavioral Anomaly Detection). Алгоритмічний аналіз семантичної густини. Замість того, щоб шукати сліди AI у синтаксисі, сучасні системи захисту повинні виявляти аномалії у поведінкових

[779] Yousefi Y., McIntyre A., Russo F., Sanchez Galera M. D., Tonello T. Regulatory innovations and policy options for generative AI and digital democracy GANs : Deliverable D6.2. *SOLARIS Project (HORIZON Grant Agreement 101094665)*. 2026. 48 p. C. 29. URL: https://projects.illc.uva.nl/solaris

[780] Rostova E. V. Psychological Surface Vectors: Mitigating Large Language Model-Driven Social Engineering via Behavioral Anomaly Detection. *The American Journal of Engineering and Technology*. 2025. Vol. 7, № 10. P. 178–184. C. 178. URL: https://www.theamericanjournals.com/index.php/tajet/article/view/6944

[781] Oriyano S. Deceptive Intelligence: AI, Social Engineering, and Securing the Human Element. New York : Apress, 2026. 727 p. C. 230. ISBN 979-8-8688-2176-9

векторах. Дослідження Ростової доводить високу ефективність гібридних моделей, що поєднують мережі довгої короткострокової пам'яті (LSTM) для аналізу послідовностей та алгоритми «Ізолюючого лісу» (Isolation Forests) для оцінки аномалій. Така архітектура аналізує відхилення в семантичній щільності та патернах комунікації. У ході симуляцій така система поведінкового виявлення успішно виявила 88,5 % витончених LLM-атак (порівняно з 34 % традиційних фільтрів)[782].

Проблема «ідеальної пам'яті» AI (AI's Perfect Memory Problem). Оріяно пропонує використовувати для верифікації характерні слабкості самих генеративних систем. Однією з таких слабкостей є «ідеальна фотографічна пам'ять» AI[783]. У той час як людині властиво забувати дрібні деталі, робити паузи, виявляти дрейф тональності та контекстуальну непослідовність, персистентний чат-бот чи цифровий аватар демонструє неприродну узгодженість упродовж тривалих періодів часу. Використання систем потокової аналітики (наприклад Apache Kafka) дозволяє відстежувати ці сеанси зв'язку в реальному часі. Якщо комунікація демонструє нелюдську стабільність (нульовий дрейф особистості та бездоганне повернення до минулого контексту), система позначає її як високоймовірну AI-децепцію[784].

Інфраструктура нульової довіри: Криптографічна та Out-of-Band верифікація. Зважаючи на те, що, як ми бачимо на практиці, людське сприйняття тепер легко вводиться в оману, а машинні детектори та фільтри можуть бути тепер скомпрометовані з великою ймовірністю, освітня корпоративна та соціальна безпека вимагає нового підходу – трансформації до криптографічної аутентифікації намірів. Перший аспект такого переходу реалізується криптографічними стандартами походження контенту (C2PA). Галузь кібербезпеки поступово переходить від спроб виявлення дипфейків (detection) до підтвердження справжності (provenance). Такі стандарти, як специфікація Coalition for Content Provenance and Authenticity (C2PA), передбачають вбудовування криптографічних метаданих про походження та зміни в медіафайли прямо на етапі їх створення (наприклад, на апаратному рівні камери або мікрофона). На відміну від традиційних водяних знаків (watermarking), які можуть бути видаленими зловмисниками (analog-hole attacks), C2PA забезпечує

[782] Rostova E. V. Psychological Surface Vectors: Mitigating Large Language Model-Driven Social Engineering via Behavioral Anomaly Detection. *The American Journal of Engineering and Technology*. 2025. Vol. 7, № 10. P. 178–184. C. 178. URL: https://www.theamericanjournals.com/index.php/tajet/article/view/6944

[783] Oriyano S. Deceptive Intelligence: AI, Social Engineering, and Securing the Human Element. New York : Apress, 2026. 727 p. C. 515–516. ISBN 979-8-8688-2176-9

[784] Там само. С. 515.

концепцію «захисту в глибину» (defense-in-depth), доводячи криптографічну цілісність цифрового сліду[785].

Наступний надійний спосіб – позасмугова автентифікація (Out-of-Band Authentication). Для захисту критичних бізнес-процесів (наприклад, фінансових транзакцій чи зміни конфігурації обладнання чи коду) верифікація через комунікаційні канали (голос, відео) має бути визнана недійсною за замовчанням. Необхідний перехід до позасмугової автентифікації (Out-of-Band Authentication, OOB). OOB спирається на володіння фізичним пристроєм та організаційними знаннями, які AI не може синтезувати віддалено.

Якщо «синтетичний керівник» (deepfake) віддає терміновий наказ через відеозв'язок, протокол безпеки повинен вимагати криптографічного підпису (через смарт-карти, YubiKey або розподілені системи на кшталт KeePassXC) у незалежному каналі (наприклад, захищений американський месенджер Signal), прив'язаному до апаратного токену. Використання криптографічних підписів (GPG/OpenPGP) в рутинне листування гарантує, що навіть якщо зловмисник згенерував ідеальний текст, він не зможе підробити математичний доказ особистості відправника[786].

Найважливішим аспектом безпеки є те, як людина-оператор взаємодіє з захисними AI-алгоритмами. Ідея включення «людини в контур» (human-in-the-loop) покликана компенсувати технічні помилки машин. Проте емпіричні дослідження свідчать про складну психологічну динаміку цього процесу. У масштабному дослідженні Локонте та співавторів вивчалося, як люди реагують на вердикти AI-детектора брехні, що володіє різним ступенем точності (54% проти 89%) і виражає різний ступінь впевненості у своїх передбаченнях (від «нерішучого» до «дуже впевненого»)[787]. Експеримент із використанням 373 учасників та кривих помилок (ROC/AUC) призвів до парадоксального висновку: люди схильні відхиляти та ігнорувати передбачення високоточного AI (89%), якщо AI демонструє високу впевненість у тому, що аналізований текст є брехнею (accusatory judgments). Тобто особливість результату дослідження проявилася тим, що чим вище була впевненість AI у тому, що людина бреше, тим сильніше

[785] Yousefi Y., McIntyre A., Russo F., Sanchez Galera M. D., Tonello T. Regulatory innovations and policy options for generative AI and digital democracy GANs : Deliverable D6.2. *SOLARIS Project (HORIZON Grant Agreement 101094665)*. 2026. 48 p. С. 29–30. URL: https://projects.illc.uva.nl/solaris

[786] Oriyano S. Deceptive Intelligence: AI, Social Engineering, and Securing the Human Element. New York : Apress, 2026. 727 p. С. 491. ISBN 979-8-8688-2176-9

[787] Loconte R., Monaro M., Pietrini P., Verschuere B., Kleinberg B. Humans incorrectly reject confident accusatory AI judgments. *Computers in Human Behavior*. 2026. Vol. 182. С. 1, 8. DOI: https://doi.org/10.1016/j.chb.2026.108530

оператори-люди коригували цей вердикт у бік «правди», штучно занижуючи ефективність детектування. Ще одна особливість гібридного прийняття рішень полягає в тому, що люди еволюційно схильні вірити комунікації та припускати чесність за умовчанням[788]. Цей парадокс пояснюється когнітивною «Теорією правдоподібності за замовчанням» (Truth-Default Theory, TDT). Крім того, звинувачення іншої людини у брехні несе високі соціальні витрати (social costs). Тому коли алгоритм впевнено класифікує повідомлення як обман, оператори відчувають психологічний дискомфорт і відхиляють вердикт машини. Додатковим фактором є те, що люди схильні систематично переоцінювати власні здібності у розпізнаванні брехні. (Dunning-Kruger effect), вважаючи себе компетентнішими, ніж алгоритм[789].

Це означає, що впровадження систем AI-виявлення в контур кібербезпеки не дасть повноцінного очікуваного ефекту, якщо оператори (SOC-аналітики, модератори, служба безпеки) не будуть обізнані про власні когнітивні спотворення і не навчаться алгоритмічної недовіри.

Регуляторні інновації, AI Act та Холістичний підхід. На макрорівні безпека від AI-децепції та феномену «Слопаганди» (Slopaganda) – алгоритмічної повені інформаційного простору низькоякісним маніпулятивним контентом – потребують суттєвих інституційних та регуляторних змін. Розглянемо обмеження Європейського закону про AI та «Парадокс обмеженого ризику». Дослідники проекту SOLARIS зазначають, що поточні законодавчі рамки, включаючи AI Act[790], страждають від «Парадоксу обмеженого ризику» (Limited-Risk Paradox)[791]. Відповідно до AI Act, генеративні AI, що створюють дипфейки, віднесені до категорії систем обмеженого ризику. Від них вимагається лише базове маркування та прозорість (повідомити користувача, що він взаємодіє з AI). Однак дослідники вважають цей захід глибоко помилковим[792] з двох причин:

1. Як вже було показано раніше, прозорість сама по собі не покращує розпізнавання (perception-accuracy gap) і може спричинити «штраф

[788] Loconte R., Monaro M., Pietrini P., Verschuere B., Kleinberg B. Humans incorrectly reject confident accusatory AI judgments. *Computers in Human Behavior*. 2026. Vol. 182. C. 2. DOI: https://doi.org/10.1016/j.chb.2026.108530

[789] Там само. С. 17.

[790] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending several legislative acts (Artificial Intelligence Act). *Official Journal of the European Union*. 2024. L Series. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[791] Yousefi Y., McIntyre A., Russo F., Sanchez Galera M. D., Tonello T. Regulatory innovations and policy options for generative AI and digital democracy GANs : Deliverable D6.2. *SOLARIS Project (HORIZON Grant Agreement 101094665)*. 2026. 48 p. C. 26. URL: https://projects.illc.uva.nl/solaris

[792] Там само. С. 27.

за прозорість» (Transparency Penalty), при якому маркування контенту як згенерованого AI знижує довіру до решти достовірного контенту.

2. Дипфейки та «мікрофейки» (гіперлокалізований синтетичний контент, націлений на вузькі аудиторії чи конкретних осіб) здатні підірвати демократичні процеси, паралізувати роботу компаній або знищити публічну репутацію (image-based sabuse), що за рівнем втрат ніяк не відповідає статусу «обмеженого ризику».

Для ефективної селекції загроз експерти SOLARIS пропонують спиратися на Акторно-мережну теорію (Actor-Network Theory, ANT) та генеративну семіотику. Дипфейк слід розглядати не як ізольований технічний артефакт, а як складний соціотехнічний конструкт, який набуває своєї руйнівної сили тільки всередині мережі репостів, алгоритмів рекомендацій, відсутності фактчекінгу та вразливостей інтерфейсу[793].

Отже, регуляторні інновації повинні змістити фокус з «поліцейського контролю за правдою» (policing truth), який технологічно недосяжний через гонку озброєнь у сфері AI і загрожує безальтернативною цензурою, на стратегії підвищення цифрової грамотності, інституційної координації та оцінки рівня шкоди (harm-based assessments).

Регуляторам необхідно вимагати від розробників (upstream) створення моделей, безпечних за замовчанням (Safe by Design), від платформ (midstream) – модифікації алгоритмів рекомендацій для маркування непідтверджених синтетичних медіа, а від держави (downstream) – інвестицій у програми AI-грамотності (AI literacy), щоб перебудувати патерни довіри громадян. Більше того, необхідні чіткі роз'яснення, сувора криміналізація та оперативне видалення глибоких фейків, що завдають непропорційної репутаційної та психологічної шкоди вразливим групам населення.

5 Проектування загроз та стратегій захисту цифрових систем підтримки навчального процесу у світлі «когнітивних конфліктів»

Інтеграція великих мовних моделей (LLM) та генеративного штучного інтелекту (AI) у критичні інфраструктури, включаючи академічне середовище (academia) та системи підтримки навчального процесу (LMS, Ed-Tech), відкрила безпрецедентні можливості для персоналізації

[793] Yousefi Y., McIntyre A., Russo F., Sanchez Galera M. D., Tonello T. Regulatory innovations and policy options for generative AI and digital democracy GANs : Deliverable D6.2. *SOLARIS Project (HORIZON Grant Agreement 101094665)*. 2026. 48 p. C. 15–30. URL: https://projects.illc.uva.nl/solaris

освіти[794]. Однак ці технології автоматично перетворили цифровий освітній простір на арену «когнітивної війни». Як зазначає Оріяно, у цій новій ері «полем битви є не далека країна чи захищена мережа. Кожна швидкоплинна взаємодія, яка здається надто реальною, щоб ставити її під сумнів, знаходиться у вашій поштовій скриньці або у вашій стрічці новин»[795].

Цифрові системи навчального процесу (університетські портали, системи дистанційного тестування, платформи e-learning) стають уразливими не лише для класичних кібератак, але й для індустріалізованого обману (industrialized deception). Студенти, викладачі та адміністративний персонал зазнають мікротаргетованих атак, що експлуатують їх психологічний стан, когнітивну втому та академічний стрес.

Розглянемо психологічну поверхню атаки (Psychological Surface Vectors) у освітньому процесі. Освітній процес характеризується високою циклічністю та періодами пікового когнітивного навантаження (сесії, дедлайни захист курсових робіт, періоди зарахування). Зловмисники використовують AI для картування цих ритмів та експлуатації психологічної поверхні атаки. Сучасні AI-інструменти (такі як Chronos Framework або Temporal Engine) здатні аналізувати хронобіологічні ритми та патерни активності студентів та викладачів. AI обчислює «вікна когнітивної вразливості», коли критичне мислення знижено через втому. Атака може виглядати як термінове повідомлення від деканату або служби IT-підтримки університету про необхідність «негайно підтвердити облікові дані для доступу до порталу», надіслане у той момент, коли студент або викладач максимально виснажений. LLM здатні автоматично отримувати психологічні профілі (включаючи модель «Великої п'ятірки») з відкритих даних (OSINT), таких як студентські форуми, соціальні мережі та академічні публікації[796]. Згенеровані AI фішингові листи або повідомлення в університетських форумах ідеально підлаштовуються під страхи або інтереси жертви (наприклад, загроза відрахування або пропозиція престижного стажування), досягаючи рівня успішності (click-through rate), більш ніж на 40 %, що перевищує традиційний фішинг[797].

---

[794] Schiffman N., Ardis J. Bridging Human Deception to AI Exploitation: A Systematic Review of Psychological Strategies in LLM Model Manipulation. *Preprints.org*. 2026. С. 6. DOI: https://doi.org/10.20944/preprints202605.0075.v1

[795] Oriyano S. Deceptive Intelligence: AI, Social Engineering, and Securing the Human Element. New York : Apress, 2026. 727 p. С. 4. ISBN 979-8-8688-2176-9

[796] Rostova E. V. Psychological Surface Vectors: Mitigating Large Language Model-Driven Social Engineering via Behavioral Anomaly Detection. *The American Journal of Engineering and Technology*. 2025. Vol. 7, № 10. P. 178–184. С. 178. URL: https://www.theamericanjournals.com/index.php/tajet/article/view/6944

[797] Там само.

Окремою нішевою проблемою є загрози синтетичних медіа та контенту в академічному середовищі. Критична вразливість застосування LLM як освітніх помічників (tutoring systems) полягає в їх здатності генерувати переконливу дезінформацію. Дослідження показують, що впровадження тонких неточностей у багатоходові діалоги з AI-репетиторами може призводити до формування хибних спогадів (false memories) у студентів. Цілком виразно можна констатувати, що серія таких атак зможе призвести до серйозної репутаційної шкоди для цифрової платформи ведення навчальної діяльності. Слопатанда як метод підриву академічного дискурсу. Генеративний AI дозволяє зловмисникам алгоритмічно «заповняти» (flood the zone) студентські форуми, платформи для дискусій та соціальні мережі низькоякісним маніпулятивним контентом – слопагандою[798]. Ця тактика використовує ботів зі збереженням багатосеансового контексту (multi-session state modeling), які безшовно імітують поведінку реальних студентів, поширюючи дезінформацію про політику університету, результати наукових досліджень чи розклад. Згодом це викликає «епістемічну фрагментацію», коли претенденти можуть втрачати здатність відрізняти факти від алгоритмічного шуму. У системах цифрової підтримки освітніх ініціатив величезну роль відіграють відео та аудіокомунікації. Технології (такі, як Coqui XTTS або FaceSwap) дозволяють створювати дипфейки учасників освітнього процесу в реальному часі[799]. Атакуючий алгоритм може ініціювати фальшивий відеодзвінок у Zoom/Teams/Meet, вимагаючи небезпечних та несанкціонованих дій від респондента, використовуючи авторитет та штучну терміновість і точно вивірений AI-асистентами момент.

Дослідники виділяють окрему категорію загроз усередині самих цифрових систем – поява емерджентних «темних патернів», згенерованих AI[800]. У цифрових освітніх системах (особливо популярних курсах та MOOC-платформах) це проявляється через: виявлення параметрів штучного дефіциту (Artificial Scarcity): сканування адаптивних таймерів, що прискорюються, коли система фіксує сумніви студента при оплаті курсу або здачі тесту. Ще одним патерном є приховані цикли

[798] Klincewicz M., Alfano M., Fard A. E. Slopaganda: The interaction between propaganda and generative AI. *Filosofiska Notiser*. 2025. Årg. 12, Nr. 1. P. 135–162. C. 138, 147. DOI: https://doi.org/10.48550/arXiv.2503.01560

[799] Yousefi Y., McIntyre A., Russo F., Sanchez Galera M. D., Tonello T. Regulatory innovations and policy options for generative AI and digital democracy GANs : Deliverable D6.2. *SOLARIS Project (HORIZON Grant Agreement 101094665)*. 2026. 48 p. C. 18. URL: https://projects.illc.uva.nl/solaris

[800] Pandey D. Emergent Dark Patterns in AI-Generated User Interfaces : arXiv preprint arXiv:2602.18445. Mathura : GLA University, 2026. 15 p. C. 8. DOI: https://doi.org/10.48550/arXiv.2602.18445

підкріплення (Reinforcement Loops): динамічне переміщення елементів інтерфейсу (наприклад, кнопки відмови від платної підписки на додаткові навчальні матеріали), що базується на аналізі DOM-мутацій та затримок введення користувача.

Інтеграція AI в системи прокторингу (контролю за іспитами) та перевірки на плагіат породжує критичні вразливості на рівні прийняття рішень людиною. Парадокс присутності генеративного AI в освітніх платформах проявляється з особливою силою. Коли студентам стає зрозумілим, що частина навчального матеріалу або зворотного зв'язку була згенерована AI, довіра до платформи різко падає – виникає «штраф за прозорість» (Transparency Penalty)[801]. Це створює парадокс: системи підтримки навчання можуть стати неефективними, оскільки будь-яке розкриття використання AI веде до системної ерозії довіри до всіх навчальних матеріалів.

Захист освітніх систем за доби «когнітивних війн» не може спиратися виключно на технічні фільтри. Необхідний підхід, що поєднує поведінковий аналіз, криптографію, медіаграмотність та регуляторні політики.

Традиційні системи виявлення вторгнень (IDS) неефективні проти AI-фішингу. Цифрові освітні платформи мають бути переведені на системи UEBA (User and Entity Behavior Analytics)[802]. Платформи e-learning повинні аналізувати семантичну щільність повідомлень студентів та співробітників, а також вимірювати ентропію взаємодії. Якщо «студент» або «викладач» демонструє ідеальну алгоритмічну пам'ять, неприродну узгодженість або нетипову для нього швидкість реакції (AI Perfect Memory Problem), система повинна позначати це як аномалію.

Для протидії дипфейкам та підробці академічних досягнень університети мають відмовитися від візуальної та аудіальної довіри. Будь-які критичні дії (зміна оцінок, видача дипломів, доступ до даних досліджень) повинні підкріплюватись стандартами доказу походження (наприклад, специфікаціями C2PA). Більше того, необхідна позасмугова аутентифікація (Out-of-Band Authentication) – наприклад, використання апаратних ключів FIDO2 або токенів, які AI-агент не здатний перехопити віддалено, навіть згенерувавши ідеальний голос або поведінку респондента.

У рамках європейського Плану дій з цифрової освіти (Digital Education Action Plan) концепція медіаграмотності має бути розширена до AI Literacy

[801] Loth A., Kappes M., Pahl M.-O. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)*, April 13–17, 2026, Dubai, United Arab Emirates. New York : ACM, 2026. P. 1–8. C. 2, 4. DOI: https://doi.org/10.1145/3774905.3795471

[802] Oriyano S. Deceptive Intelligence: AI, Social Engineering, and Securing the Human Element. New York : Apress, 2026. 727 p. C. 442. ISBN 979-8-8688-2176-9

(AI-грамотності)[803]. Для навчання студентів та співробітників необхідно використовувати метод «пребункінгу» (prebunking) та гейміфікацію. Прикладом є тренінги “Bad News” і “Harmony Square”, де учасники самі приміряють роль творців дезінформації, щоб зрозуміти механіку маніпуляцій[804]. Крім того, для адміністраторів LMS та IT-персоналу повинні проводитися регулярні мікро-тренування (Micro-Drills) зі стрес-інокуляцією (Stress Inoculation), де симулюються когнітивно складні атаки (псевдо-термінові дзвінки, комбіновані з фішинговими повідомленнями в месенджері), щоб виробити навичку обов’язкової криптографічної перевірки навіть під тиском.

Згідно з Акторно-мережевою теорією (ANT), дипфейки та AI-обман набувають чинності тільки всередині соціально-технічної мережі. Керівництво освітніх установ має впроваджувати політику «Безпеки за умовчанням» (Safe by Design) і використовувати локально-розгорнуті моделі (LLaMA, Ollama), доступні лише через суворі політики управління (OPA/Rego), щоб запобігти витоку дослідницьких даних та унеможливити ін’єкції навчальних вибірок для ботів.

Висновки

Інформаційна безпека в епоху генеративного AI вимагає фундаментальних трансформацій. Злиття машинного навчання та когнітивної психології породило алгоритмічних агентів, здатних обходити традиційні технічні засоби захисту (IDS, фільтри) шляхом експлуатації людських слабкостей (довіри, поспіху, емпатії) через персистентну багатосеансову взаємодію. Захист інформаційної цілісності сьогодні неможливий шляхом простого вдосконалення детекторів дипфейків, оскільки «гонка озброєнь» завжди залишає детектор на крок за генератором. Справжня безпека та верифікація в умовах AI-децепції вимагають нових піджодів. Перехід до поведінкового аналізу аномалій (дрейф тональності, ентропія, ідеальна пам’ять AI), що реалізується через моделі машинного навчання (LSTM, Isolation Forests). Впровадження криптографічних стандартів (C2PA) та суворої позасмугової апаратної верифікації (Out-of-band authentication),

[803] Yousefi Y., McIntyre A., Russo F., Sanchez Galera M. D., Tonello T. Regulatory innovations and policy options for generative AI and digital democracy GANs : Deliverable D6.2. *SOLARIS Project (HORIZON Grant Agreement 101094665)*. 2026. 48 p. C. 40. URL: https://projects.illc.uva.nl/solaris

[804] Klincewicz M., Alfano M., Fard A. E. Slopaganda: The interaction between propaganda and generative AI. *Filosofiska Notiser*. 2025. Årg. 12, Nr. 1. P. 135–162. C. 151. DOI: https://doi.org/10.48550/arXiv.2503.01560

що усувають суб'єктивну візуальну або аудіальну довіру. Навчання операторів алгоритмічній недовірі для подолання когнітивних спотворень (Truth-Default Theory), які змушують людей помилково відкидати об'єктивні обвинувальні висновки AI. Адаптації регуляторної бази (розширення AI Act) для визнання дипфейків та мікрофейків системами високого ризику та перекладу фокусу з цензури контенту на криптографічне підтвердження його походження та системне підвищення алгоритмічної грамотності (AI literacy) суспільства.

У рамках побудови та обслуговування інформаційних освітніх систем на даному етапі також з'являється низка першочергових завдань. Система повинна захищати претендентів від фішингу та маніпуляцій, які використовують «когнітивні спотворення» (bias). Необхідно впроваджувати протоколи багатофакторної аутентифікації (MFA) та криптографічних підписів для верифікації навчальних матеріалів, розпоряджень адміністрації, конфігураційних та сервісних дій на платформах. Обов'язковою умовою експлуатації стає застосування AI для виявлення аномалій у діях користувачів, що дозволяє виявляти спроби захоплення облікових записів, ін'єкцій коду та інших компрометуючих цілісність системи дій. Одночасно з цим необхідно дотримуватися правил етики та конфіденційності даних. Збір докладної телеметрії про дії студента (час реакції, патерни введення, особливості сеансу роботи) має бути суворо регламентований відповідно до відповідних стандартів захисту даних (GDPR/HIPAA) та прозорості алгоритмів аналізу поведінки.

Лише глибока системна інтеграція поведінкових, криптографічних, когнітивних та правових стратегій захисту здатна забезпечити інституційну стійкість та протистояти індустріалізованій автоматизації цифрового обману.

Цифрові системи підтримки навчального процесу опинилися на передовій у когнітивних війнах. Атакуючі алгоритми можуть використовувати потужність LLM, генеративних агентів та глибокої психометрики для експлуатації вразливостей учасників освітнього процесу, минаючи традиційні мережеві екрани. Індустріалізована машина цифрового обману вражає за найважливішим активом академічної установи – довірою та об'єктивністю сприйняття.

Комплексний захист освітніх закладів вимагає переходу до інфраструктури нульової довіри до контенту. Це передбачає повсюдне використання поведінкової аналітики аномалій, криптографічних сертифікатів автентичності (C2PA), позасмугової (Out-of-Band) перевірки критичних академічних рішень і, що найважливіше, безперервної стрес-інокуляції

та розвитку просунутої алгоритмічної грамотності серед усіх учасників навчального процесу для збереження системної безпеки та демократичних цінностей цифрової освіти.

Анотація

Мета публікації полягає у теоретико-методологічному обґрунтуванні та розробці комплексних стратегій і соціотехнічних архітектур проактивного захисту цифрових систем підтримки освітнього процесу, спрямованих на нейтралізацію кібер-когнітивних загроз штучного інтелекту, збереження когнітивного суверенітету та забезпечення стратегічної стабільності інституцій в умовах гібридного протистояння.

В основі роботи лежить комплексний системний та соціотехнічний підхід, опис методів цифрової психології обману, теорії машинного навчання та поведінкового аналізу. Для реалізації систем детекції аномалій запропоновані гібридні алгоритмічні моделі, що поєднують нейронні мережі довгої та короткострокової пам'яті (LSTM) та алгоритми ізолюючого лісу (Isolation Forests). Верифікація контенту базується на інтеграції сучасних криптографічних стандартів походження інформації та інструментів потокової аналітики реального часу.

З позицій соціотехнічного аналізу систематизовано та класифіковано новітні кібер-когнітивні загрози для цифрових освітніх екосистем, зокрема вектори психологічної поверхні (*Psychological Surface Vectors*), які експлуатують індивідуальні особливості користувачів за моделлю «Великої п'ятірки» (*Big Five*) за допомогою великих мовних моделей (LLM).

Детально досліджено психологічну динаміку взаємодії людини та машинних інтерфейсів, зокрема когнітивний «розрив між сприйняттям та фактичною точністю» (Perception-Accuracy Gap), за якого суб'єктивна довіра користувача до емпатичних та впевнених AI-агентів переважає над об'єктивною істинністю наданих ними суджень.

Запропоновано використання методу ідентифікації AI-децепції за допомогою моніторингу «проблеми ідеальної пам'яті» штучного інтелекту (AI's Perfect Memory Problem) у реальному часі через системи потокової аналітики, що базується на фіксації неприродної довгострокової контекстуальної узгодженості та нульового дрейфу особистості цифрового агента.

Обґрунтовано модель інфраструктури нульової довіри (*Zero Trust Infrastructure*) для освітніх мереж, що передбачає впровадження криптографічних стандартів підтвердження походження контенту (C2PA) та механізмів *Out-of-Band* верифікації намірів.

Література

1. Schiffman N., Ardis J. Bridging Human Deception to AI Exploitation: A Systematic Review of Psychological Strategies in LLM Model Manipulation. *Preprints.org*. 2026. DOI: https://doi.org/10.20944/preprints202605.0075.v1
2. Pandey D. Emergent Dark Patterns in AI-Generated User Interfaces : arXiv preprint arXiv:2602.18445. Mathura : GLA University, 2026. 15 p. DOI: https://doi.org/10.48550/arXiv.2602.18445
3. Loth A., Kappes M., Pahl M.-O. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)*, April 13–17, 2026, Dubai, United Arab Emirates. New York : ACM, 2026. P. 1–8. DOI: https://doi.org/10.1145/3774905.3795471
4. Rostova E. V. Psychological Surface Vectors: Mitigating Large Language Model-Driven Social Engineering via Behavioral Anomaly Detection. *The American Journal of Engineering and Technology*. 2025. Vol. 7, № 10. P. 178–184. URL: https://www.theamericanjournals.com/index.php/tajet/article/view/6944 (дата звернення: 23.05.2026).
5. Klincewicz M., Alfano M., Fard A. E. Slopaganda: The interaction between propaganda and generative AI. *Filosofiska Notiser*. 2025. Årg. 12, Nr. 1. P. 135–162. DOI: https://doi.org/10.48550/arXiv.2503.01560
6. Yousefi Y., McIntyre A., Russo F., Sanchez Galera M. D., Tonello T. Regulatory innovations and policy options for generative AI and digital democracy GANs : Deliverable D6.2. *SOLARIS Project (HORIZON Grant Agreement 101094665)*. 2026. 48 p. URL: https://projects.illc.uva.nl/solaris (дата звернення: 25.05.2026).
7. Loconte R., Monaro M., Pietrini P., Verschuere B., Kleinberg B. Humans incorrectly reject confident accusatory AI judgments. *Computers in Human Behavior*. 2026. Vol. 182. DOI: https://doi.org/10.1016/j.chb.2026.108530
8. Cole S. New Study Reveals the Manipulative "Dark Patterns" of AI Chatbots. *404 Media*. 2026. URL: https://www.404media.co/new-study-reveals-the-manipulative-dark-patterns-of-ai-chatbots (дата звернення: 25.05.2026).
9. Oriyano S. Deceptive Intelligence: AI, Social Engineering, and Securing the Human Element. New York : Apress, 2026. 727 p. ISBN 979-8-8688-2176-9
10. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending several legislative acts (Artificial Intelligence Act). *Official Journal of the European Union*. 2024. L Series. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (дата звернення: 25.05.2026).