УДК 004.056:004.8

Лобанов Іван Григорович,

студент І курсу Харківського національного автомобільно-дорожнього університету

Діденко Наталя Вікторівна,

к.т.н., доцент кафедри кібербезпеки Харківського національного автомобільно-дорожнього університету

ІНТЕЛЕКТУАЛЬНІ МЕТОДИ ВИЯВЛЕННЯ ТА ПРОТИДІЇ СОЦІАЛЬНІЙ ІНЖЕНЕРІЇ НА ОСНОВІ ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ

Анотація. У роботі проаналізовано вплив генеративного штучного інтелекту на еволюцію атак соціальної інженерії та обґрунтовано підходи до їх виявлення і протидії в корпоративних інформаційних системах. Розглянуто три основні напрями підсилення загроз: створення реалістичного контенту, персоналізоване таргетування та автоматизація інфраструктури атак. Запропоновано гібридну модель захисту, що поєднує традиційні заходи підвищення обізнаності персоналу з інтелектуальним моніторингом комунікацій на основі великих мовних моделей. Підхід орієнтовано на інтеграцію з рамкою NIST Cybersecurity Framework 2.0 та враховує специфіку гібридних загроз. Результати вказують на доцільність переходу від реактивних до проактивних стратегій кіберстійкості.

Ключові слова: соціальна інженерія, генеративний штучний інтелект, кіберстійкість, великі мовні моделі, менеджмент інформаційної безпеки.

У сучасних умовах цифровізації та зростання гібридних загроз соціальна інженерія залишається одним із найбільш ефективних векторів кібератак, оскільки експлуатує людський фактор, а не технічні вразливості [1, 2]. Традиційні технічні засоби захисту (міжмережеві екрани, системи виявлення вторгнень, антивірусне програмне забезпечення) часто виявляються недостатніми, коли атакуючий отримує доступ через довірчі відносини з співробітниками організації. Поява генеративних моделей штучного інтелекту, зокрема великих мовних моделей, суттєво змінила ландшафт таких атак, зробивши їх більш масштабованими, персоналізованими та переконливими [1].

Як показано в роботі Schmitt і Flechais, генеративний штучний інтелект підсилює соціальну інженерію за трьома ключовими напрямами: створення реалістичного контенту, що імітує людську комунікацію; розширене таргетування та персоналізація на основі відкритих даних; автоматизація інфраструктури атак, що дозволяє одночасно проводити кампанії великого масштабу [1]. Це знижує бар'єр входу для атакуючих і підвищує ймовірність успіху навіть проти підготовлених організацій. Дослідження Kumarage та співавторів демонструє можливість моделювання багатоходових атак соціальної інженерії за допомогою агентних систем на основі великих мовних моделей, а також підходи до їх виявлення з урахуванням психологічних профілів жертв [2].

Аналіз еволюції загроз, представлений у огляді Yu та співавторів, виділяє три фази розвитку AI-підсиленої соціальної інженерії: розширення масштабу через цифрові медіа, збагачення новими векторами атак та появу якісно нових методів [3]. У цьому контексті класичні підходи до навчання персоналу, описані ще в класичній роботі Mitnick і Simon, потребують доповнення технологічними рішеннями реального часу [4]. Особливо актуальною стає проблема в умовах кібервійни та гібридних операцій, коли атакуючі можуть поєднувати соціальну інженерію з іншими методами впливу на критичну інфраструктуру та системи цивільного захисту.

Для протидії таким загрозам доцільно застосовувати гібридний підхід, що інтегрує організаційні, освітні та технологічні заходи. З організаційної точки зору важливим є впровадження елементів NIST Cybersecurity Framework 2.0, зокрема функцій Govern, Protect, Detect та Respond, з акцентом на категорії обізнаності та навчання персоналу, а також моніторингу аномальної активності [5]. Технологічний компонент передбачає використання великих мовних моделей для аналізу комунікаційних потоків (електронна пошта, месенджери, голосові канали) з метою виявлення ознак соціальної інженерії: нетипових запитів на

конфіденційні дані, тиску часом, створення штучної терміновості, підробки ідентичності авторитетних осіб.

Ефективність таких систем може бути підвищена за рахунок персоналізації захисту. Моделі, що враховують поведінкові патерни та психологічні характеристики співробітників, здатні краще розрізняти легітимні та шкідливі взаємодії [2]. Водночас необхідно забезпечувати баланс між точністю виявлення та збереженням приватності, уникаючи надмірного нагляду. Важливим елементом є також регулярне оновлення навчальних програм з урахуванням нових можливостей генеративних моделей, включаючи deepfake-аудіо та відео, які вже використовуються в реальних інцидентах.

Практична реалізація запропонованого підходу передбачає поетапне впровадження: спочатку аудит поточних процесів управління ризиками та рівня обізнаності персоналу; далі – інтеграція модулів інтелектуального аналізу комунікацій у існуючі SOC або SIEM-системи; паралельно – розробка сценаріїв навчання, що імітують сучасні AI-генеровані атаки. Такий підхід дозволяє не лише реагувати на інциденти, але й підвищувати загальну кіберстійкість організації. Особливу увагу слід приділяти секторам, пов’язаним із цивільним захистом та критичною інфраструктурою, де наслідки успішної соціальної інженерії можуть мати системний характер.

Дослідження Ai, Kumarage та співавторів (2024) підкреслює двоїсту природу великих мовних моделей у контексті чат-орієнтованої соціальної інженерії (CSE). Моделі здатні генерувати високоякісний шкідливий контент, але їхні можливості виявлення атак без додаткової настройки залишаються обмеженими. Автори запропонували модульний пайплайн ConvoSentinel, який аналізує повідомлення на рівні запитів чутливої інформації, фрагментів розмови та повної діалогової сесії з використанням retrieval-augmented generation. На датасеті SEConvo система досягає macro

F1 на рівні 0.80 при використанні GPT-4-Turbo, що перевищує базові few-shot підходи, і демонструє ефективне раннє виявлення вже після п’яти повідомлень.

Ці результати підтверджують доцільність використання гібридних архітектур, де спеціалізовані детектори на рівні окремих повідомлень доповнюються агентними системами для оцінки загального контексту розмови. У корпоративному середовищі такі підходи можуть бути інтегровані в існуючі процеси моніторингу електронної пошти та внутрішніх комунікацій без необхідності повного перегляду інфраструктури. Важливо, що ConvoSentinel знижує операційні витрати порівняно з чистими few-shot методами за рахунок меншої кількості токенів у запитах до моделей.

Окремої уваги заслуговує проблема deepfake-контенту в соціальній інженерії. Хоча класичні текстові атаки залишаються домінуючими, поява доступних інструментів генерації голосу та відео створює додаткові вектори, зокрема в сценаріях вішингу та відеодзвінків. Дослідження вказують на необхідність багаторівневого захисту, що включає як технічне виявлення синтетичних медіа, так і організаційні процедури верифікації ідентичності співрозмовника. У контексті українських організацій, які функціонують в умовах підвищених кіберризиків, такі заходи мають бути частиною загальної стратегії кіберстійкості.

З точки зору менеджменту інформаційної безпеки, впровадження інтелектуальних методів протидії соціальній інженерії вимагає оновлення політик і процедур. Зокрема, рекомендується включити в програми підвищення обізнаності сценарії, згенеровані за допомогою великих мовних моделей, щоб персонал міг розпізнавати сучасні тактики. Паралельно слід забезпечити можливість ескалації підозрілих комунікацій до фахівців SOC без створення надмірного навантаження на співробітників. Інтеграція з

NIST CSF 2.0 дозволяє структурувати ці заходи в межах функцій Protect (обізнаність і навчання) та Detect (моніторинг аномалій).

Обмеження поточного стану досліджень полягають у тому, що більшість експериментів проводяться на синтетичних датасетах. Реальні корпоративні середовища мають вищу варіативність комунікаційних стилів, культурних особливостей та внутрішніх процедур. Тому пілотне впровадження гібридних систем доцільно супроводжувати збором зворотного зв'язку від користувачів і постійною валідацією моделей на внутрішніх даних (з дотриманням вимог конфіденційності). Це дозволяє уникнути високого рівня хибних спрацювань, які можуть знизити довіру персоналу до системи захисту.

Таким чином, генеративний штучний інтелект одночасно є джерелом нових загроз і інструментом їх нейтралізації. Перехід до інтелектуальних методів виявлення та протидії соціальній інженерії, інтегрованих із сучасними рамками управління кіберризиками, є необхідною умовою забезпечення кіберстійкості в умовах швидкої еволюції атак. Подальші дослідження доцільно спрямувати на розробку відкритих наборів даних для навчання моделей виявлення AI-генерованих атак соціальної інженерії та на оцінювання ефективності гібридних систем захисту в реальних організаційних середовищах. Особливу увагу слід приділити адаптації таких рішень до потреб організацій критичної інфраструктури та системи цивільного захисту України.

Список використаних джерел

1. Schmitt M., Flechais I. Digital deception: generative artificial intelligence in social engineering and phishing. Artificial Intelligence Review. 2024. Vol. 57. Art. no. 324. DOI: https://doi.org/10.1007/s10462-024-10973-2.

2. Kumarage T., Johnson C., Adams J., Ai L., Kirchner M., Hoogs A., Garland J., Hirschberg J., Basharat A., Liu H. Personalized Attacks of Social Engineering

in Multi-turn Conversations: LLM Agents for Simulation and Detection. arXiv preprint arXiv:2503.15552. 2025. DOI: https://doi.org/10.48550/arXiv.2503.15552.

3. Yu J., Yu Y., Wang X., Lin Y., Yang M., Qiao Y., Wang F.-Y. The Shadow of Fraud: The Emerging Danger of AI-powered Social Engineering and its Possible Cure. arXiv preprint arXiv:2407.15912. 2024. DOI: https://doi.org/10.48550/arXiv.2407.15912.

4. Mitnick K. D., Simon W. L. The Art of Deception: Controlling the Human Element of Security. New York : Wiley, 2002. 368 p.

5. National Institute of Standards and Technology. The NIST Cybersecurity Framework (CSF) 2.0. NIST CSWP 29. 2024. DOI: https://doi.org/10.6028/NIST.CSWP.29.