Бобер Н.В., магістрантка
Воротніков В.В., д.т.н, професор
Державний університет «Житомирська політехніка»

АНАЛІЗ ВИКОРИСТАННЯ ChatGPT В СОЦІАЛЬНІЙ ІНЖЕНЕРІЇ: НОВІ ПІДХОДИ ДО КІБЕРЗАГРОЗ

У сучасному цифровому середовищі значення кібербезпеки різко зросло, ставши ключовим аспектом організаційної стратегії та національної безпеки. Ця ескалація насамперед пов'язана зі зростаючою залежністю від інформаційно-комунікаційних технологій у різних секторах.

Однією з найбільш значущих нових технологій, що впливають на кібербезпеку, є штучний інтелект (ШІ). Роль штучного інтелекту в кібербезпеці подвійна: він не тільки розширює можливості систем кібербезпеки, але й представляє нові вразливості, якими можуть скористатися кіберзлочинці.[1]

Ландшафт кібербезпеки, що розвивається, став свідком грізного супротивника у вигляді атак соціальної інженерії. Оскільки технології продовжують розвиватися, зловмисники соціальної інженерії все частіше використовують складні інструменти та методи, щоб обдурити своїх цілей. Одним з таких інструментів, який останніми роками привернув значну увагу, є генеративний штучний інтелект. [2]

ШІ відкриває нову еру в царині соціальної інженерії, надаючи суб'єктам загрози та кіберзлочинцям передові інструменти та тактики маніпулювання, обману та компроментації комп'ютерних систем. Ці технологічні досягнення відкривають зловмисникам кілька можливостей використовувати штучний інтелект для оркестрування атак соціальної інженерії.

В рамках дослідження розроблено: метод тестування здатності ChatGPT генерувати контент для атак соціальної інженерії; систему оцінки рівня довіри користувачів до згенерованих текстів; алгоритм аналізу лінгвістичних особливостей шкідливих повідомлень; методику перевірки вразливостей існуючих антифішингових систем перед контентом, створеним ШІ.

Запропонована система дослідження включає три основні компоненти:

- генерацію тестових даних, що створює різні типи шкідливих текстів (фішингові листи, маніпулятивні повідомлення, сценарії телефонного шахрайства);
- аналітичний модуль, що виконує лінгвістичний аналіз контенту, оцінюючи його маніпулятивні характеристики та схожість із реальними шахрайськими схемами;
- підсистему тестування безпеки, яка перевіряє ефективність антифішингових механізмів у виявленні контенту, створеного ChatGPT.

Попередні експерименти та теоретичний аналіз розробленої системи дослідження дозволяє прогнозувати високу ефективність її застосування для виявлення атак соціальної інженерії. Очікується, що система дослідження буде здатна:

- класифікувати згенеровані ChatGPT фішингові повідомлення;
- виявляти лінгвістичні патерни, характерні для маніпулятивних атак;
- аналізувати ефективність існуючих антифішингових систем та виявляти їх слабкі місця.

Для підтвердження ефективності запропонованого підходу планується проведення серії експериментів, у яких користувачам буде запропоновано оцінити достовірність згенерованих повідомлень [2], а також тестування зразків на сучасних антифішингових алгоритмах.

Попередні результати демонструють, що ChatGPT здатний генерувати контент, який може обійти деякі механізми захисту, що вказує на необхідність удосконалення методів кібербезпеки.

Практична цінність дослідження полягає у можливості вдосконалення існуючих антифішингових систем через врахування особливостей текстів, створених ШІ, дозволить підвищити рівень інформаційної безпеки та зменшити ризик атак соціальної інженерії.

Список використаних джерел:

1. Artificial Intelligence's Impact on Social Engineering Attacks. URL: https://opus.govst.edu/cgi/viewcontent.cgi?article=1521&context=capstones (дата звернення 24.02.2025)
2. Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks. URL: https://www.researchgate.net/publication/374536568_Decoding_the_Threat_Landscape_ChatGPT_FraudGPT_and_WormGPT_in_Social_Engineering_Attacks (дата звернення 24.02.2025)
3. The Impact of Using ChatGPT on Cybersecurity & Social Engineering. URL:
4. https://www.researchpublish.com/upload/book/The%20Impact%20of%20Using%20ChatGPT-16112023-4.pdf (дата звернення 24.02.2025)