УДК 630,247
О. Оробчук, Ph.D; А. Мазяр
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

«АНАЛІЗ ВРАЗЛИВОСТЕЙ ШТУЧНОГО ІНТЕЛЕКТУ ТА МЕТОДІВ ЇХНЬОГО ВИКОРИСТАННЯ В КІБЕРЗАГРОЗАХ»

UDC 630,247
O. Orobchuk, Ph.D; A. Maziar

«ANALYSIS OF ARTIFICIAL INTELLIGENCE VULNERABILITIES AND METHODS OF THEIR USE IN CYBER THREATS»

Стрімкий розвиток штучного інтелекту (ШІ) відкриває нові горизонти для інновацій, проте разом із цим створює потенційні ризики кібербезпеки. Зловмисники активно досліджують і використовують вразливості ШІ для реалізації складних кіберзагроз, що вимагає поглибленого вивчення цієї проблематики.

У сучасному середовищі штучного інтелекту існує кілька типів вразливостей, які можуть використовуватися зловмисниками. Перш за все, варто зазначити, що одним із поширених видів атак є вплив на вхідні дані. Це може проявлятися у вигляді отруєння даних, коли до системи вводяться некоректні або навмисно шкідливі дані, що порушують її роботу, або створення спеціальних шкідливих прикладів, які вводять моделі в оману. Додатково, моделі ШІ самі можуть мати внутрішні слабкі сторони, як-от перенавчання чи нездатність коректно працювати з неточними чи некоректними даними. Ще однією серйозною загрозою є використання моделей для викрадення конфіденційної інформації, що може реалізуватися через атаки інверсії моделей.

Такі вразливості вже активно використовуються у реальних кіберзагрозах. Генеративні моделі, наприклад, можуть допомагати у проведенні фішингових кампаній та здійсненні атак із використанням соціальної інженерії. Ще одним небезпечним застосуванням є створення deepfake-контенту, що використовується для маніпуляцій і поширення дезінформації. Окрім того, автономні системи, зокрема дрони та транспортні засоби, також мають свої вразливі місця, які можуть бути використані зловмисниками для створення загроз.

Таким чином, проблема вразливостей ШІ потребує системного підходу для забезпечення кібербезпеки. Запропоновані рекомендації можуть бути інтегровані у сучасні стратегії безпеки.

Література

1. Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Practical black-box attacks against machine learning. Proceedings of the ACM on Artificial Intelligence.
2. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition.
3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.