УДК 004.7:519.87(043.3):621.391 DOI: 10.31673/2412-4338.2023.033140

Бойко А.О., Шулімова Д.Д.
Державний університет інформаційно-комунікаційних технологій, Київ

МОДЕЛЬ ЗАГРОЗ СИСТЕМ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ, ЩО ФУНКЦІОНУЮТЬ НА ОСНОВІ ЕЛЕМЕНТІВ ШТУЧНОГО ІНТЕЛЕКТУ

Анотація. *Сучасний розвиток штучного інтелекту (ШІ) впливає на багато аспектів життя і діяльності суспільства, включаючи системи підтримки прийняття рішень. ШІ вже успішно впроваджується у різних сферах, таких як бізнес, медицина, автоматизована обробка даних та багато інших. Проте, цей розвиток також приносить з собою різноманітні загрози та виклики, які потребують уважного вивчення та заходів безпеки. Дана стаття присвячена розробці та аналізу моделі загроз для систем підтримки прийняття рішень, які функціонують на основі елементів штучного інтелекту (ШІ). Актуальність дослідження полягає у швидкому розвитку технологій ШІ та їх все більш широкому використанні в різних галузях, включаючи управління, медицину, фінанси та багато інших. У статті описується структура та основні складові моделі загроз, що можуть виникнути в системах підтримки прийняття рішень на основі ШІ. Наведено огляд існуючих атак на штучний інтелект (ШІ), які охоплюють етапи навчання, використання алгоритмів машинного навчання, а також інформаційну інфраструктуру системи. Виконано класифікацію даних атак, засновану на аналізі процесу функціонування ШІ, та виявлено категорії атак, найбільш актуальні та небезпечні для систем ШІ. Запропоновано модель загроз для системи підтримки прийняття рішень, заснованої на технологіях штучного інтелекту, особливістю якої є врахування загроз конфіденційності даних ШІ систем, загроз функціонуванню ШІ систем та загроз інформаційним системам. Дана модель має практичне значення для розробників систем ШІ та для фахівців у галузі кібербезпеки, оскільки воно сприяє підвищенню рівня захищеності та надійності систем підтримки прийняття рішень на основі елементів штучного інтелекту у сучасному цифровому середовищі.*

Ключові слова: *штучний інтелект, машинне навчання, нейронні мережі, людино-машинна взаємодія, візуалізація даних, комп'ютерна безпека, уразливості технологій*

Boiko A., Shulimova D.
State University of Information and Communication Technologies, Kyiv

THREAT MODEL OF DECISION SUPPORT SYSTEMS OPERATING ON THE BASIS OF ELEMENTS OF ARTIFICIAL INTELLIGENCE

Abstract*. The modern development of artificial intelligence (AI) affects many aspects of society's life and activities, including decision support systems. AI is already successfully implemented in various fields such as business, medicine, automated data processing and many others. However, this development also brings with it various threats and challenges that require careful study and security measures. This article is devoted to the development and analysis of a threat model for decision support systems that function on the basis of elements of artificial intelligence (AI). The relevance of the research lies in the rapid development of AI technologies and their increasingly widespread use in various fields, including management, medicine, finance and many others. The article describes the structure and main component models of threats that may*

arise in AI-based decision support systems. An overview of existing attacks on artificial intelligence (AI) is given, which cover the stages of training, the use of machine learning algorithms, as well as the information infrastructure of the system. Classification of attack data was carried out, based on the analysis of the AI functioning process, and attack categories most relevant and dangerous for AI systems were identified. A threat model for a decision support system based on artificial intelligence technologies is proposed, the feature of which is the consideration of threats to the confidentiality of data of AI systems, threats to the functioning of AI systems, and threats to information systems. This model is of practical importance for developers of AI systems and for specialists in the field of cyber security, as it contributes to increasing the level of security and reliability of decision support systems based on artificial intelligence elements in the modern digital environment.

Keywords: *artificial intelligence, machine learning, neural networks, human-machine interaction, data visualization, computer security, technology vulnerabilities*

1. Вступ

Розвиток сучасних інформаційних технологій призводить до того, що технології машинного навчання з використанням нейронних мереж стають все більш популярними. Однак ця область залишається маловивченою. Щодо нової галузі комп'ютерної безпеки є інформаційна безпека моделей штучного інтелекту. У зв'язку з цим актуальним стає завдання розробки спеціалізованої моделі загроз систем штучного інтелекту.

Ця модель враховує уразливості нейронної мережі, вхідні дані на етапі її навчання, а також можливе спотворення даних у процесі функціонування комп'ютерних систем. Крім того, вона аналізує можливі загрози інформаційній безпеці та включає розроблену класифікацію загроз технологіям штучного інтелекту. Ця модель дозволяє розуміти потенційні загрози технологіям штучного інтелекту, що може бути корисним для розробників нейронних мереж та додатків, що базуються на елементах машинного навчання, а також для дослідників, які займаються безпекою таких технологій.

2. Сучасний рівень розвитку штучного інтелекту

Сучасні інформаційні технології поступово інтегруються у структуру бізнесу та широко впроваджуються у конкретних додатках [1]. Незважаючи на те, що технологія штучного інтелекту (ШІ) з'явилася нещодавно, вона стрімко розвивається і перед розробниками постійно постають нові завдання. Розглянемо деякі результати, одержані з оглядів, де обговорювалися проблеми машинного навчання.

Технологія подвійного призначення ШІ може бути як благословенням, так і прокляттям для кібербезпеки [2]. Це підтверджується тим фактом, що ШІ використовується як для підтримки шкідливих атак, так і для протистояння ризикам кібербезпеки.

Атака на ШІ є цілеспрямоване маніпулювання системою ШІ з кінцевою метою викликати її збій.

Змагальне машинне навчання (англ. Generative Adversarial Learning, GAN) - це підход у сфері машинного навчання, в якому дві нейронні мережі, відомі як "генератор" і "дискримінатор", змагаються одна з одною в процесі навчання.

Головною ідеєю GAN є створення моделі (генератора), яка здатна генерувати дані, які схожі на реальні дані з навчального набору. Ця генерована інформація піддається оцінці і "критиці" (дискримінатору), який намагається відрізнити ці синтетичні дані від реальних. Генератор намагається постійно покращувати якість своїх вихідних даних так, щоб дискримінатор не міг їх відрізнити від реальних. [3].

У порівнянні з традиційними програмно-апаратними системами, системи на основі ШІ мають специфічні функції, які можуть атакуватися нетрадиційними способами. Зокрема, набір навчальних даних може бути компрометований таким чином, що "навчання" системи не відповідатиме задуманому. Можливість вибору зовнішніх об'єктів, які виявлятимуться системою, може бути змінена так, що система не зможе їх розпізнати. Тому важливо

забезпечити додатковий спеціалізований захист систем ШІ, щоб гарантувати їх безпечний життєвий цикл розробки від ідеї до розгортання та спостереження після виходу на ринок, включаючи моніторинг та аудит під час виконання.

Оскільки системи ШІ інтегровані в критично важливі комерційні та військові програми, атаки на них можуть мати серйозні наслідки. Атаки з використанням ШІ можуть бути використані різними способами для досягнення зловмисних кінцевих цілей.

Крім того, через загальне застосування елементів ШІ в повсякденному житті людини, крім впровадження нових і потужних напрямків для проведення атак на вразливості систем ШІ, самі ШІ будуть все частіше використовуватися для здійснення атак на подібні системи.

3. Класифікація загроз технологіям штучного інтелекту

Атаки на моделі штучного інтелекту (ШІ) часто класифікуються за трьома основними напрямками: вплив на класифікатор, порушення безпеки та специфіка цих атак. Вони також можуть бути додатково поділені на "білу скриньку" та "чорну скриньку". У разі атак "білої скриньки" зловмисник має доступ до параметрів моделі. У разі атак "чорної скриньки" такого доступу у зловмисника немає [4].

Проблема у тому, що зловмисник може порушити процес навчання моделей ШІ, запровадивши у них троянського коня. Заразив моделі троянською атакою, поведінка системи ШІ може стати непередбачуваною. Один із способів перевірки моделей ШІ на надійність - це використання так званої троянської програми, тобто шкідливої програми, яка виконує несанкціоновані користувачем дії. Троянська програма змінює модель ШІ, яка на той час навчається реагувати на хибні імпульсні сигнали введення. Ці сигнали призводять до виведення неправильної відповіді. Для більш відтворюваних і масштабованих тестів дослідники з Університету Джона Гопкінса розробили платформу під назвою TrojAI [4], яка є набором інструментів, що створюють набори даних і пов'язані моделі ШІ з троянами. Вони стверджують, що це дозволить зрозуміти вплив різних конфігурацій наборів даних на створені троянські моделі та допоможе всебічно тестувати нові методи виявлення троянських програм захисту моделей.

Атаки на вхідні дані призводять до збоїв у роботі системи ШІ, змінюючи вхідні дані, що надходять до системи. Це досягається шляхом впровадження шаблону атаки на вхідні дані, наприклад, шляхом впровадження мітки або внесення невеликих змін у цифрове зображення, що завантажується у соціальну мережу. В умовах звичайного використання система ШІ приймає допустимі вхідні дані, обробляє їх за допомогою моделі (мозку) та повертає вихідні дані. Під час атаки введені вхідні дані до системи ШІ змінюються з використанням шаблону атаки, внаслідок чого система ШІ повертає неправильний висновок [4].

Атаки на вхідні дані не вимагають, щоб зловмисник пошкодив систему ШІ для її атаки. Навіть дуже сучасні системи ШІ, що відрізняються високою точністю і ніколи не порушують своєї цілісності, залишаються вразливими для атак введення. Важливо, що, на відміну інших кібератак, сама атака який завжди використовує комп'ютер.

Атаки ухилення є найбільш поширеним типом атак, під час яких зловмисник намагається уникнути виявлення своїх дій, заплутуючи та модифікуючи вміст листів та шкідливих програм. Ухилення не передбачає впливу на дані, що використовуються для навчання моделі. Прикладом атаки ухилення є спам на основі зображень, де спам вбудований в прикріплене зображення, щоб уникнути аналізу моделями машинного навчання, призначеними для захисту від спаму. Іншим прикладом є атаки спуфінгу на системи біометричної автентифікації на основі ШІ [2-5].

Зараження - ще один тип атаки, що є "вороже зараження" даних. Системи машинного навчання часто перенавчаються на даних, зібраних у процесі їх роботи, і зловмисник може "заразити" ці дані, впровадивши шкідливі зразки, які пізніше порушать процес перенавчання. Зловмисник може вводити дані на етапі навчання, які помилково позначені як безпечні, хоча насправді вони є шкідливими. Наприклад, великі мовні моделі, такі як OpenAI GPT-3,

можуть розкривати конфіденційну особисту інформацію під час введення певних слів та фраз.

У той час як крадіжка моделі, яка також називається вилученням моделі, передбачає, що зловмисник досліджує "чорну скриньку" системи машинного навчання, щоб або відновити модель, або витягти дані, на яких вона була навчена. Це може викликати проблеми, якщо навчальні дані чи сама модель є конфіденційними. Наприклад, крадіжка моделі може використовуватися для отримання конфіденційної торгової моделі акцій, яку зловмисник може використовувати для отримання фінансової вигоди.

4. Модель загроз

Розглянемо основні компоненти процесу функціонування системи штучного інтелекту (ШІ):

1. Учасник: Це окремі особи, підприємства або організації, які надають дані для навчання системи ШІ.
2. Серверна частина: Це апаратно-програмна платформа, на якій функціонує система ШІ.
3. Розробник: Розробник відповідає за обробку вихідних даних, створення алгоритму навчання і взаємодію з користувачем.
4. Користувач: Користувач взаємодіє з системою ШІ для отримання кінцевого результату роботи.

Процес функціонування системи ШІ можна поділити на два основні етапи:

1. Етап навчання: На цьому етапі здійснюється збір і обробка вихідних даних, а також сам процес навчання системи. Учасники навчання надають дані для навчання, а розробник реалізує модель і алгоритм навчання.
2. Етап використання: Це безпосереднє використання системи ШІ в реальному середовищі. На цьому етапі взаємодіє лише користувач системи.

Кожен з цих етапів має свій набір даних та функціоналу, що може бути цінним для зловмисників. На рис. 1 наведено схему цих етапів роботи системи ШІ та її компонентів.

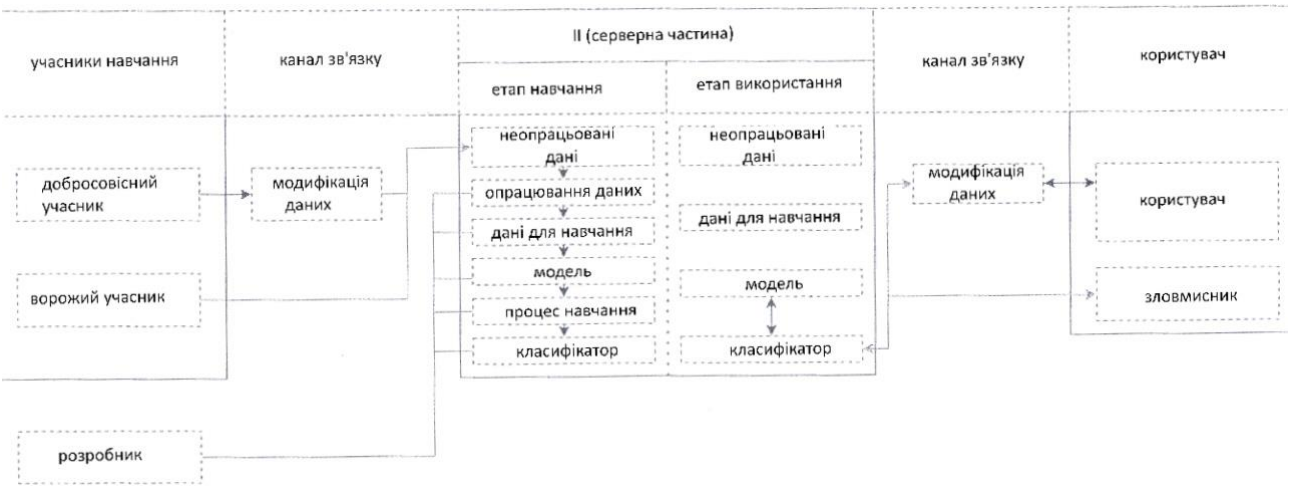

Рис. 1 – Етапи роботи системи штучного інтелекту

Розглянемо кожен етап:

1. Етап навчання: Цей етап є основним і найдовшим, відповідаючи за коректність роботи системи ШІ. Одним з головних фокусів на цьому етапі є дані для навчання. Система ШІ, навчена на різних наборах даних, може показувати різні результати. Велика частина атак зловмисників спрямована на спотворення даних для зниження рівня навчання. Також на цьому етапі можуть відбуватися атаки розробників, які можуть внести "лазівки" під час реалізації класифікатора або модифікувати дані для навчання [6].

2. Етап використання: Цей етап відбувається в промисловому середовищі і включає взаємодію користувача з системою ШІ. На цьому етапі вся взаємодія відбувається з користувачем системи ШІ (з винятком "лазівок" розробників).

На етапі використання системи ШІ зловмисникам можуть бути цікаві такі цілі:

— Дані, що використовуються в моделі: Зловмисники можуть намагатися отримати доступ до даних, які використовуються системою ШІ для прийняття рішень.

— Дані, що використовувалися в навчанні: Зловмисники можуть намагатися отримати доступ до даних, на основі яких система ШІ навчалася, що може розкрити вразливості чи обмеження системи.

— Злам алгоритму роботи системи ШІ: Зловмисники можуть намагатися підірвати або змінити алгоритм роботи системи ШІ, щоб спотворити результати.

— Дані, які надають користувачі: Зловмисники можуть спробувати отримати доступ до введених користувачами даних для використання в своїх цілях.

Таблиця 1

Класифікація атак на системи ШІ

№ п/п	Класифікаційна ознака комп'ютерної атаки	Індекс класифікатора	Зміст класифікатора
1	По активності ШІ	S1	На етапі навчання
			На етапі користування
2	По суб'єкту дії	S2	На користувача системи ШІ
			На учасника системи ШІ
			На систему ШІ
			На розробника системи ШІ
3	По об'єкту атаки	S3	Модель
			Класифікатор
			Дані для навчання
4	За обізнаністю зловмисника	S4	«білий ящик»
			«сірий ящик»
			«чорний ящик»
5	Цілові/нецільові	S5	Цільові
			Нецільові
6	Обмеження методу	S6	I-0
			I-1
			I-2
			I-безкінечність
7	Ціль атаки	S7	Шпигунство
			Саботаж
			Шахрайство
8	Атакований алгоритм машинного навчання	S8	Класифікація
			Регрес
			Генеративні моделі
			Кластеризація
			Зменшення розмірності
			Навчання

9	Джерело атаки	S9	Зовнішня
			Внутрішня

З перерахованого вище можна зробити висновок, що система ШІ може бути схильною до загроз у різних елементах (див. рис. 1), на різних етапах використання та від різних учасників. Пропоновану класифікацію атак на системи ШІ можна знайти в табл. 1.

Розглянемо типи цих атак:

1. S1 - Класифікація атак заснована на етапах роботи ШІ [1,7, 8]: Ця класифікація вказує на те, на якому етапі роботи ШІ відбувається атака. Вона включає такі типи атак:

Спотворення розмітки: В атаках цього типу зловмисники додають об'єкти з неправильною розміткою до навчальної вибірки, що може призвести до неправильного навчання алгоритму [6,9];

Спотворення навчальної вибірки: Зловмисники додають спеціальні об'єкти до навчальної вибірки, які погіршують якість роботи алгоритму [10,11];

Ухилення: Це найпоширеніший тип атак, де дані змінюються для уникнення виявлення або класифікуються як законні [2].

2. S2 - класифікація атак, заснована на суб'єкті, який зазнає шкоди. Наведемо приклади можливих наслідків для кожної сторони взаємодії ШІ:

— можливим збитком для користувача системи ШІ буде: втрата персональних, біометричних даних, втрата матеріальних цінностей або шкода здоров'ю;

— виведення з ладу є основним і катастрофічним збитком системи ШІ;

— учасник навчання, аналогічно користувачеві, схильний до крадіжки персональних, біометричних даних;

— навчена система ШІ і набір даних для навчання мають велику цінність для розробника, їхнє викрадення і подальше використання зловмисником можуть завдати великої шкоди розробнику.

3.S3 - класифікація атак, заснована на об'єкті атаки в системі ШІ, що становить інтерес для зловмисника. Можна виділити такі об'єкти, які становлять важливість:

— дані для навчання - оскільки система ШІ, навчена на різних даних, має різну якість роботи, а навчальна вибірка може містити персональні або біометричні дані, виникає необхідність захисту даних для навчання;

— модель даних, яка є своєрідним "ключем" для проведення атак, таких як атаки на класифікатор і реалізація власного ШІ (конкуренція);

— класифікатор - основна мета зловмисників, спрямована на обман, виведення з ладу реалізованих алгоритмів ШІ [5].

4.S4 - класифікація атак, заснована на рівні обізнаності зловмисника:

— метод "чорної скриньки" - зловмисник може тільки надіслати інформацію в систему й отримати простий результат про клас;

— метод "сірої скриньки" - зловмисник може знати детальну інформацію про набір даних або тип нейронної мережі, її структуру, кількість шарів;

— метод "білої скриньки" - усе про мережу відомо, включно з усіма вагами і всіма даними, на яких цю мережу було навчено [2,9,12].

5. S5 - класифікація атак, заснована на характері дії зловмисника. Одним з основних завдань системи ШІ є класифікація об'єктів і залежно від мети зловмисника (приховати або замаскувати предмет) атаки можна поділити на такі види:

— цільові - атаки, спрямовані на маскування одного об'єкта під інший (класичним прикладом цієї атаки є злом біометричної ідентифікації [13]);

— нецільові - атаки, спрямовані на приховування об'єктів під час розпізнавання (приховування номерів машин для камер дорожнього руху [14,15]). S6 - Класифікація атак

заснована на рівні спотворення вхідних даних: Ця класифікація вказує на те, які зміни в вхідних даних можуть призвести до невірних результатів в роботі алгоритмів систем ШІ.

S7 - Класифікація атак, заснована на характері дії зловмисника і переслідуваних цілях: В цій класифікації атаки розділяються залежно від характеру дій зловмисника та їх цілей, такі як шпигунство, саботаж і шахрайство.

S8 - Класифікація атак, ґрунтована на розв'язуваних завданнях системи ШІ: Ця класифікація вказує на типи атак, які використовуються для розв'язування конкретних завдань ШІ, таких як класифікація, генеративні моделі, регресія, кластеризація тощо.

S9 - Класифікація атак, заснована на місцезнаходженні загрози:

— внутрішні - у вихідному коді системи ШІ можуть знаходити лазівки, що дадуть змогу здійснювати атаки (одним із таких уразливих місць для здійснення атаки є передача даних між шарами нейронної мережі);

— зовнішні - загрози, що виходять ззовні системи; до прикладів таких атак можна віднести такі втручання:

1) висновок про приналежність - атака, спрямована на приналежність даних до вибірки для навчання;

2) запам'ятовування моделі/вилучення моделі - атаки, спрямовані на викрадення моделі даних [3,4,9].

Звернемо свою увагу на ключові аспекти безпеки в роботі систем штучного інтелекту (ШІ). Важливою є інфраструктура, включаючи програмно-апаратну платформу, мережеву взаємодію, розмежування рівнів доступу та логування, які сприяють захисту системи від потенційних загроз.

Серед ключових елементів для забезпечення надійності ШІ є правильна класифікація атак та їх аналіз. У табл. 2 наведено класифікацію атак на системи ШІ, що ілюструє приклади різних типів атак і їхніх можливих наслідків для систем ШІ. Це допомагає більш глибоко зрозуміти сценарії загроз і їх вплив на функціонування системи.

Таблиця 2

Класифікація атак на інформаційні системи ШІ

№ п/п	Класифікаційна ознака комп’ютерної атаки	Індекс класифікатора	Зміст класифікатора
1	Джерело атаки	D1	Внутрішній
			Зовнішній
			Підміна даних під час мережевої взаємодії (атака «людина посередині»)
2	Частина інфраструктури що атакується	D2	Міжмережева взаємодія
			Клієнтська частина
			Серверна частина
3	Потенційний збиток	D3	Низький
			Середній
			Високий
4	За обізнаністю зловмисника	D4	Навмисне
			Ненавмисне
5	Можливість проведення	D5	Низька
			Середня
			Висока

У табл. 3 наведено приклади атак, націлених на конфіденційність даних систем ШІ, які демонструють такі відомості про загрозу:

1.етап роботи системи ШІ;

2.жертва атаки;

3.атакуючий.

Таблиця 3

Загрози конфіденційності даних систем ШІ

		Атакуючий				
		Етап навчання			Етап використання	
		Учасники	Сервер	Розробник	Сервер	Користувач
Жертва	Учасники	Витік даних учасника			Висновок про належність/запам’ятовування моделі	
	Сервер					Вилучення моделі/ухилення
	Розробник				Вилучення моделі	Вилучення моделі
	Користувач				Порушення конфіденційності користувача	

У табл. 4 наведено приклади ураження функціонування систем ШІ, що демонструють аналогічні відомості.

Суттєво, що виявлено зв'язок між надійністю системи ШІ та якістю вихідних даних для навчання. Ця залежність підкреслює важливість забезпечення достовірності та цілісності даних на етапі навчання, що впливає на відповідність роботи системи її завданням.

Таблиця 4

Загрози функціонуванню систем ШІ

		Атакуючий				
		Етап навчання			Етап використання	
		Учасники	Сервер	Розробник	Сервер	Користувач
Жертва	Учасники					
	Сервер					Вилучення моделі/ухилення
	Розробник				Вилучення моделі	Вилучення моделі
	Користувач	Отруєння				
		Лазівки у системі ШІ				

Детальна розбивка загроз на кілька класифікацій дозволяє більш глибоко аналізувати вразливості та шукати можливі шляхи їх подолання. Перевірка вихідного коду і системного середовища може виявити лазівки і захистити систему від вразливостей та потенційної атаки.

Як видно з табл. 3 і 4, що уважне дослідження можливих наслідків атак на системи ШІ допомагає зрозуміти потенційні ризики та підходи до їх запобігання. Детальна класифікація атак і їх наслідків створює можливість для розробки ефективних стратегій безпеки для систем штучного інтелекту.

6. Висновок

У статті наведено приклади атак на системи ІШ на основі аналізу яких запропоновано ознаки класифікації з урахуванням різних етапів, елементів та учасників. Розглянуто додаткову класифікацію атак на інформаційні системи. Наведено приклади класифікації атак за кількома категоріями з демонстрацією результатів аналізу того, які загрози є актуальними, а які - ні.

Список використаної літератури

1. Clark,G.A Malicious Attack on the Machine Learning Policy/G.Clark, M.Doran, W.Glisson//17th IEEE International Conference On Trust, Security And Privacy in Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering(TrustCom/BigDataSE), 2018.

2. Adversarial attacks in machine learning: What they are and how to stop them. - https://venturebeat.com/2021/05/29/adversarial-attacks-in-machine-learning-what-the-are-and-how-to-stop-them.

3.Ateniese,G. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers/G.Ateniese,G.Felici,L.V.Mancini 1 et al.//International Journal of Security and Networks.-201.-Vol.10,Issue 3.-P.137-150.

4.How Adversarial Attacks Work.-https://medium.com/xix-ai/how-adversarial-attacks-work-87495b81da2d.

5.Artificial intelligence and cybersecucity.-https://www.techopedia.com/artificial-intelligence-in-cybersecurity/2/34390.

6.Springer,J.M. A little Robustness Goes a Long Way Leveraging Robust features for Targeted Transfer Attacks/J.M.Springer, M.Mitchell, G.T.Kenyon//35th Conference on Neural Information Processing Systems (NeurIPS,2021).

7.Pang, T. Accumulative Posoning Attacks on Real-time Data/T.Pang,X.Yang, Y.Dong et al.//35th Conference on Neural Information Processing Systems (NeurIPS,2021).

8.How to attack Machine Learning (Evasion, Poisoning,Inference,Trojans,Backdoors).-https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c.

9.Poisoning attacks on Machine Learning.-https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db.

10. Liu,G. Provably Efficient Black-Box Poisoning Attacks Against Reinforcement Learning/G. Liu,L. Lai//35 Conference on Neural Information Processing Systems (NeurIPS,2021).

11.Optical Adversarial Attack Can Change Meaning of Road Signs. -https://www.unite ai/optical-adversarial-attack-can-change-the-meaning-of-road-signs/.

12.Shokri,R. Membership Inference Attack against Machine Learning Models /R.Shok, M.Stronati, C.Song, V.Shmatikov//IEEE Symposium on Security and Privacy,2017.

13.Niu,D. Moire Attack (MA): A New Potential Risk Of Screen Photos/D.Niu, R.Guo, Y.Wang//35th Conference on Neur Information Processing Systems (NeurIPS,2021).

14.Su,J. One pixel attack for fooling deep neural networks/J.Su,D.V. Vargas, S.Kouichi//IEEE Transactions on Evolutionary Computation.-2019.-Vol.23 Issue 5.-P.828-841.

References

1.Clark, G.A Malicious Attack on the Machine Learning Policy/G.Clark, M.Doran, W.Glisson//17th IEEE International Conference On Trust, Security And Privacy in Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), 2018.

2. Adversarial attacks in machine learning: What they are and how to stop them. - https://venturebeat.com/2021/05/29/adversarial-attacks-in-machine-learning-what-the-are-and-how-to-stop-them.

3. Ateniese, G. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers/G. Ateniese, G. Felici, L.V. Mancini 1 et al.//International Journal of Security and Networks.-201.-Vol.10, Issue 3 .-P.137-150.

4. How Adversarial Attacks Work.-https://medium.com/xix-ai/how-adversarial-attacks-work-87495b81da2d.

5. Artificial intelligence and cybersecurity.-https://www.techopedia.com/artificial-intelligence-in-cybersecurity/2/34390.

6. Springer, J.M. A little Robustness Goes a Long Way Leveraging Robust features for Targeted Transfer Attacks/J.M.Springer, M.Mitchell, G.T.Kenyon//35th Conference on Neural Information Processing Systems (NeurIPS, 2021).

7. Pang, T. Accumulative Posing Attacks on Real-time Data/T. Pang, X. Yang, Y. Dong et al.//35th Conference on Neural Information Processing Systems (NeurIPS, 2021).

8. How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors).-https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c.

9.Poisoning attacks on Machine Learning.-https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db.

10. Liu, G. Provably Efficient Black-Box Poisoning Attacks Against Reinforcement Learning/G. Liu, L. Lai//35 Conference on Neural Information Processing Systems (NeurIPS, 2021).

11. Optical Adversarial Attack Can Change Meaning of Road Signs. -https://www.unite ai/optical-adversarial-attack-can-change-the-meaning-of-road-signs/.

12. Shokri, R. Membership Inference Attack against Machine Learning Models /R.Shok, M.Stronati, C.Song, V.Shmatikov//IEEE Symposium on Security and Privacy, 2017.

13. Niu, D. Moire Attack (MA): A New Potential Risk Of Screen Photos/D.Niu, R.Guo, Y.Wang//35th Conference on Neur Information Processing Systems (NeurIPS,2021).

14. Su, J. One pixel attack for fooling deep neural networks/J.Su, D.V. Vargas, S. Kouichi//IEEE Transactions on Evolutionary Computation.-2019.-Vol.23 Issue 5.-P.828-841.