Issue 12, 2022

https://doi.org/10.23939/

УДК 004.89

ЗАБЕЗПЕЧЕННЯ КІБЕРБЕЗПЕКИ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ: АНАЛІЗ ВРАЗЛИВОСТЕЙ, АТАК І КОНТРЗАХОДІВ

Олексій Неретін[1], Вячеслав Харченко[2]

[1, 2] Національний аерокосмічний університет ім. М. Є. Жуковського "ХАІ",
[1] oleksii.neretin@gmail.com, ORCID 0000-0003-2114-6714,
[2] v.kharchenko@csn.khai.edu, ORCID 0000-0001-5352-077X

Останніми роками багато компаній почали інтегрувати системи штучного інтелекту (СШІ) в свої інфраструктури. СШІ використовують у вразливих сферах суспільства, таких як судова система, критична інфраструктура, відеоспостереження тощо. Це зумовлює необхідність достовірного оцінювання і гарантованого забезпечення кібербезпеки СШІ. У дослідженні проаналізовано стан справ щодо кібербезпеки цих систем. Класифіковано можливі типи атак і детально розглянуто основні з них. Проаналізовано загрози і атаки за рівнем тяжкості й оцінено ризики безпеки з використанням методу IMECA. Виявлено, що найвищі ризики небезпеки "Змагальних атак" та атак "Отруєння даних", але контрзаходи щодо них не на належному рівні. Зроблено висновок, що існує потреба в формалізації та стандартизації життєвого циклу розроблення та використання безпечних СШІ. Обґрунтовано напрями подальших досліджень щодо необхідності розроблення методів оцінювання і забезпечення кібербезпеки СШІ, зокрема для систем, які надають штучний інтелект як сервіс.

Ключові слова: штучний інтелект; кібербезпека; змагальні атаки; отруєння і витік даних; троянські атаки; атаки на модель; крадіжки и отруєння моделей; контрзаходи.

Вступ

Мотивація. Сьогодні штучний інтелект (ШІ) проникає в усі сфери життя, як і сталі технології загального призначення (інтернет, мобільні комунікації, розумні будинки тощо). Те, що раніше могли робити тільки люди, тепер передають розумним програмам і системам. ШІ відіграє все більшу роль для суспільства. Його активно застосовують в державному та фінансовому секторах, медицині, військовій справі, побуті та в інших сферах.

За прогнозами дослідницької компанії IDC, у 2022 р. світові витрати на штучний інтелект досягнуть 79,2 млрд доларів США [1]. Автори звіту PricewaterhouseCoopers [2] прогнозують, що до 2030 р. завдяки прискореному розвитку систем штучного інтелекту (СШІ) глобальний ВВП може зрости на 14 % (це орієнтовно 15,7 трлн доларів США).

Чим ширшою стає сфера використання ШІ, тим більше цим напрямом цікавляться кіберзлочинці. Сучасні СШІ ґрунтуються на методах, які вразливі щодо руйнівних атак, дуже небезпечних для їх функціонування. Завдяки цьому зловмисники можуть здобути контроль над СШІ і доволі вільно маніпулювати ними для зміни поведінки і, в кінцевому підсумку, для безпосереднього впливу на безпеку користувачів. Отже, забезпечення кіберзахисту СШІ є доволі актуальним і важливим напрямом досліджень і розробок.

Існує певний набір векторів атак на СШІ. Щоб зрозуміти становище за цим напрямом, треба класифікувати ці атаки та детально розглянути основні з них. Ґрунтуючись на класифікації,

необхідно проаналізувати атаки за рівнем небезпеки для системи загалом. Далі, на основі результатів аналізу, потрібно визначити атаки, які можуть заподіяти найбільшу шкоду і рівень контрзаходів від яких недостатній. Виявлені за допомогою такого аналізу критичні напрями стануть базовими для подальших досліджень за цією тематикою.

Аналіз джерел. Відомі джерела групують за кількома напрямами, які нижче проаналізуємо детально. В [1–10] розглянуто загальну інформацію про СШІ та стан їх кібербезпеки, подано класифікацію та опис механізмів атак за типами. Основний посил такий: ШІ може бути атакованим і на це треба зважати, розробляючи СШІ. Розглянуто як звичайні вразливості програмного забезпечення, так і специфічні вектори атак на ШІ, такі як “Атаки на платформу”, “Атаки на алгоритм” та “Атаки на дані”.

Група джерел [4, 8, 11–37] детальніше описує кожен зі специфічних типів атак на ШІ, а саме “Модифікація даних”, “Відмова в обслуговуванні”, “Вхідний витік”, “Змагальні атаки”, “Атаки з отруєнням даних”, “Витік даних” та “Атаки на модель”, які характеризуються різним рівнем ймовірності, тяжкості та, зрештою, впливу на безпеку СШІ загалом.

Джерела [6, 8, 26, 38] частково приділяють увагу захисту СШІ та аналізу вразливостей загалом, а [39–50] детальніше висвітлюють методи та рекомендації щодо забезпечення кібербезпеки ШІ. Розглянуто можливі контрзаходи для кожного типу атак та проаналізовано їх рівень.

Зауважимо, що потрібно звернути додатково увагу на питання, пов’язані із аналізом безпосереднього впливу атак на СШІ, а також рівня відповідних контрзаходів, щоб виділити найнебезпечніші та недостатньо досліджені типи кібератак.

Метою статті є аналіз найістотніших загроз, вразливостей і контрзаходів для забезпечення кібербезпеки систем штучного інтелекту, які потребують подальшого дослідження та доопрацювання.

Завдання дослідження такі:

- класифікувати атаки на ШІ за типами;
- детально проаналізувати основні атаки на СШІ за рівнем небезпечності;
- опрацювати відомі методи і рекомендації щодо забезпечення кібербезпеки ШІ;
- виявити атаки із найбільшим рівнем шкоди і низьким рівнем контрзаходів;
- обґрунтувати напрям майбутніх досліджень, зважаючи на результати аналізу.

Статтю структуровано так. Розділ 1 містить класифікацію джерел, проаналізованих під час досліджень. Розділ 2 аналізує вразливості й типи атак на СШІ. В третьому розділі оцінено ризики безпеки й ефективність контрзаходів для різних атак на СШІ із використанням методу IMECA. У четвертому розділі проаналізовано заходи міждержавної взаємодії, методи та рекомендації для забезпечення кібербезпеки СШІ. Загальні висновки за результатами аналізу і напрями подальших досліджень наведено у п’ятому і шостому розділах відповідно.

1. Класифікація джерел за напрямами

Джерела, опрацьовані для цього дослідження, поділено на сім напрямів, які надані у табл. 1, де вони також класифіковані за типами (наукові статті, звіти, інтернет-ресурси).

Таблиця 1

Класифікація джерел за напрямами

№ з/п	Напрям	Тип джерела	Джерело
1	2	3	4
1	Загальні відомості про стан кібербезпеки систем ШІ і класифікація атак	Звіт	[1, 2, 3, 5, 6, 8, 9]
		Інтернет-ресурс	[4, 7, 10]
2	Атаки на платформу (platform attacks)	Звіт	[8]
3	Змагальні атаки на ШІ (adversarial attacks)	Наукова стаття	[13, 14, 15, 16, 17, 18, 19]
		Інтернет-ресурс	[4, 11, 12]

Продовження табл. 1

1	2	3	4
4	Атаки з отруєнням даних (data poisoning attacks), троянські атаки/бекдори	Наукова стаття	[22, 23, 24, 25, 26, 27]
		Інтернет-ресурс	[20, 21]
5	Витік даних (data leakage)	Наукова стаття	[28, 29, 30, 31, 32, 33]
		Інтернет-ресурс	[34]
6	Атаки на модель (model attacks), техніка крадіжки моделей, отруєння моделей	Наукова стаття	[35, 36, 37]
7	IMECA-аналіз кібератак і контрзаходів для забезпечення безпеки СШІ	Звіт	[8, 38]

2. Огляд атак на системи штучного інтелекту

2.1. Загальні відомості про вразливості і класифікація атак на системи ШІ

Дослідження кібербезпеки СШІ виконаємо, проаналізувавши основні вразливості, атаки на них, наслідки цих атак та засоби захисту (контрзаходи).

Вразливості СШІ пов’язані із їх обмеженнями, якими успішно користуються зловмисники. Цілком достатньо ізоляційної стрічки, яка перетворить знак зупинки на зелене світло світлофора в “очах” безпілотного автомобіля [3] чи змусить його хибно вважати, що треба пришвидшуватися. Наочним прикладом є результати дослідження команди з McAfee Advanced Threat Research, завдяки якому вони змусили автопілот “Tesla” розігнатися до 85 миль за годину, замість того, щоб дотримуватися швидкості 35 миль за годину [4]. На рис. 1 надано звичайний (з лівої сторони) та модифікований (з правої) знаки обмеження швидкості. Чорна стрічка, що модифікувала цифру 3, зробивши її трохи схожою на цифру 8 (рисунок з правої сторони), змусила автопілот зі штучним інтелектом зробити критичну помилку, що могла б призвести до непередбачуваних наслідків.

Рис. 1. Звичайний та модифікований знаки обмеження швидкості (https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/)

Через різноманітні вразливості СШІ хакери мають змогу змушувати їх помилятись, надавати конфіденційну інформацію та навіть припиняти функціонування [5]. Вразливості СШІ можна розділити на дві основні групи – це “традиційні” вразливості програмного забезпечення (ПЗ) (атаки на інструментарій) та специфічні типи вразливостей, притаманні тільки цим складним системам. Виділено атаки на “традиційні” вразливості ПЗ двох основних типів [6]:

· класичні атаки на ПЗ – зазвичай це атаки на відкрите програмне забезпечення, великі кількості якого використовуються під час розроблення та функціонування СШІ. Вразливості створюють, додаючи їх у популярні продукти зі світу ШІ, а також використовуючи вже відомі на цей час вразливості цих продуктів [7];

· типосквотінг (typosquatting) – це атаки, націлені на створення бібліотек, назви яких схожі на назви популярних у галузі інструментів.

Атаки на вразливості цього типу заподіюють шкоду таким властивостям СШІ, як конфіденційність, цілісність та доступність. Атаки на специфічні вразливості мають більше різновидів. На рис. 2 подано класифікацію атак на відповідні групи вразливостей [1, 8–10].

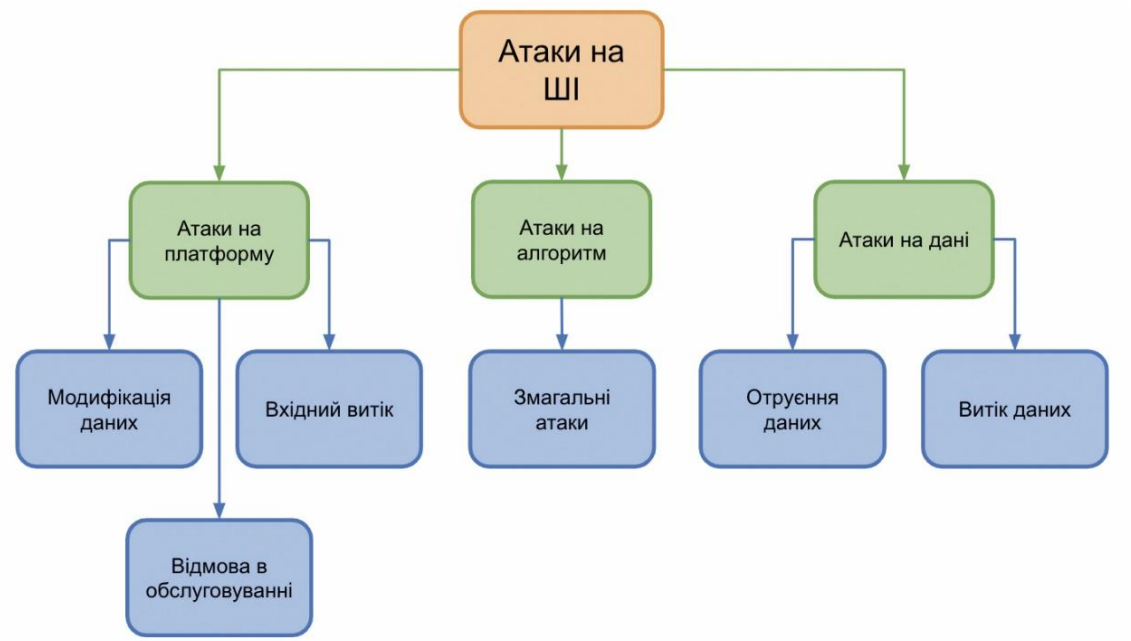

Рис. 2. Класифікація атак на системи ШІ

2.2. Атаки на платформу

Першим типом специфічних атак є “Атаки на платформу” (platform attacks), на якій працюють СШІ [8]. Є три різновиди цього типу атак:

· модифікація даних (data modification) – це маніпулювання параметрами моделі. Досягається завдяки підбору таких вхідних даних, які зможуть вивести з ладу внутрішні механізми, що обробляють ці дані. Завдяки атакам модифікації даних зловмисники впливають на цілісність СШІ;

· відмова в обслуговуванні (denial of service) – це виведення з ладу або уповільнення СШІ. Виконується за допомогою надсилання великого обсягу трафіку, який забороняє доступ звичайним користувачам або уповільнює його. Атаки цього типу безпосередньо впливають на властивість доступності системи;

· вхідний витік (input leakage) – це заволодіння вхідними даними користувачів. Відбувається за рахунок компрометації СШІ або використання вразливостей в оточенні цієї системи. Як наслідок, порушується конфіденційність даних СШІ.

2.3. Змагальні атаки на ШІ

Другим типом специфічних атак є “Атаки на алгоритм” (algorithm attacks), який використовує СШІ [8]. Основний його різновид – “Змагальні атаки на ШІ”, які ґрунтуються на додаванні “шуму” до вхідних даних, завдяки чому система робить хибне передбачення [11].

Один із прикладів цієї атаки ілюструє рис. 3, на якому зображені дослідники із Бельгійського університету KU Leuven, які зламують відеоаналітичний сервіс зі штучним інтелектом за допомогою кольорового роздрукованого патерна [12, 13]. Особу, зображену ліворуч, система легко розпізнає як людину, а особу праворуч взагалі не класифікує як людину, бо кольоровий патерн заважає їй це зробити.

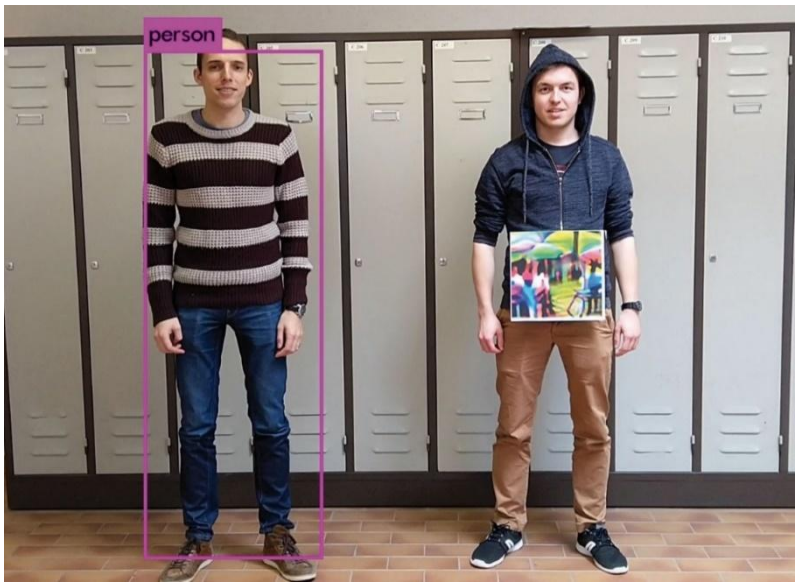

Рис. 3. Злам системи відеоаналітичного сервісу (https://www.securityinfowatch.com/video-surveillance/video-analytics/article/21080107/researchers-hack-ai-video-analytics-with-color-printout)

Форми змагальних атак можна розділити за сприйняттям на:

· видимі для людського ока шаблони атак [3, 14, 15] (на рис. 4 шматочки стрічки перетворюють знак “STOP” на зелене світло світлофора для США [3]);

· невидимі для людського ока зміни, які стають причиною хибного результату класифікування моделлю ШІ [3, 16–19] (на рис. 5 невидимий для людського ока шаблон перетворює зображення панди на мавпу для США [3]).

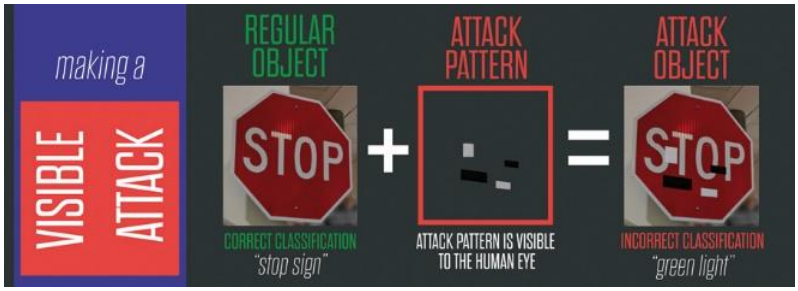

Рис. 4. Видимий для людського ока шаблон змагальної атаки (https://www.belfercenter.org/publication/AttackingAI)

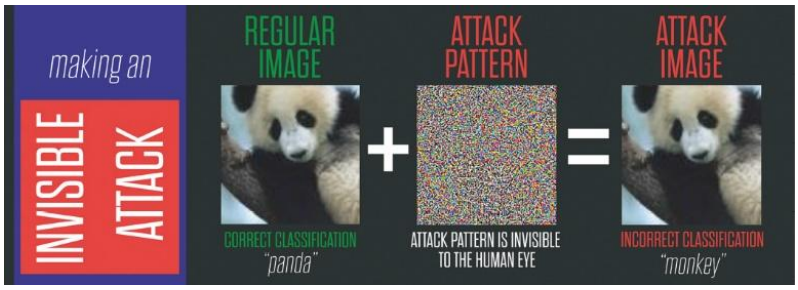

Рис. 5. Невидимий для людського ока шаблон змагальної атаки (https://www.belfercenter.org/publication/AttackingAI)

Змагальні атаки на ШІ завдають шкоду цілісності цим системам.

2.4. Атаки із отруєнням даних

Третім типом специфічних атак є “Атаки на дані” (data attacks) [8]. Є два різновиди цих атак. Перший – “Атаки з отруєнням даних” (data poisoning attacks). Відповідну вразливість експлуатують, додаючи спеціально створені екземпляри даних до навчальних наборів даних СШІ. Здебільшого ці екземпляри не заважають працездатності системи, але, передаючи певні дані на вхід до СШІ, можуть змусити її видавати хибний результат, на який і очікує зловмисник [20–27].

На рис. 6 зображені особи в чорно-білій кепці. Всіх модель хибно класифікує як Френка Сміта. Це так званий бекдор, який дає змогу в потрібний момент змусити СШІ поводитись так, як вигідно кіберзлочинцю.

Рис. 6. Отруєння моделі даних людиною у чорно-білій кепці (https://ece.duke.edu/about/news/detecting-backdoor-attacks-artificial-neural-networks)

Завдяки отруєнню даних зловмисники порушують цілісність СШІ.

Другим напрямом “Атак на дані” є “Витік даних”. Атаки витоку навчальних даних порушують конфіденційність даних СШІ. Вони націлені на визначення того, чи задіяний той чи інший набір даних у процесі навчання моделі. Також є небезпека, що зловмисник може отримати несанкціонований доступ до особистих даних користувачів системи у випадку її компрометації, якщо вона такі дані зберігає [8, 28–34].

Здійснюючи атаки витоку даних, кіберзлочинці впливають на властивість конфіденційності СШІ.

2.5. Атаки на модель

Доцільно виділити ще один напрям небезпеки для СШІ, а саме “Атаки на модель (model attacks), техніка крадіжки моделей, отруєння моделей” [9].

Цей напрям можна поділити на:

- атаки на вагові характеристики ШІ, які можуть повністю зруйнувати систему [35];
- дублювання, або повна крадіжка моделі ШІ [36];
- отруєння моделей загалом [37].

Атаки цього типу впливають на показники конфіденційності та цілісності СШІ.

3. IMECA-аналіз кібератак і контрзаходів для забезпечення безпеки СШІ

На основі класифікації атак на СШІ проаналізуємо атаки за рівнем небезпеки для систем у формальніший спосіб відповідно до основних положень методу IMECA (Intrusion Modes Effects Criticality Analysis) [38]. Щоб визначити рівень небезпеки, виконаємо аналіз за такими параметрами:

- загроза – за допомогою чого здійснюється атака на систему;
- вразливість – слабка частина системи, завдяки якій можлива атака;
- атака – тип вторгнення;
- наслідки – наслідки, які може заподіяти атака;
- ймовірність – наскільки ймовірне скоєння атаки;

- тяжкість – наскільки серйозною та небезпечною буде атака за наслідками;
- ризик – який сумарний вплив атаки на систему, що ґрунтується на ймовірності та тяжкості;
- контрзаходи – заходи та дії, спрямовані на протидію атакам.

Комбінація показників ймовірності появи та тяжкості визначає рівень ризику – показника критичності. Ефективні контрзаходи, своєю чергою, можуть зменшити критичність. Високий рівень показника тяжкості із низьким рівнем контрзаходів у тих атак, що можуть найбільше зашкодити системам ШІ. Результати IMECA аналізу наведено у табл. 2. Аналіз виконано з використанням даних [8] і він дещо узагальнений із погляду деталізації всіх елементів аналізу.

Таблиця 2

IMECA-аналіз кібератак і контрзаходів для забезпечення безпеки СШІ

#	Загроза	Вразливість	Атака	Наслідки	Критичність			Контр-заходи
					Ймовір-ність	Тяж-кість	Ризик	
1	2	3	4	5	6	7	8	9
1	Ретельно підібрані вхідні дані, які можуть модифіку-вати параметри СШІ	Сторонній код, недостатня перевірка вхідних даних	Моди-фікація даних	Втрата цілісності системи	Середня	Низька	Низький	Відмова від зайвого сторон-нього коду; ретельний контроль коду, який працює із вхідними даними
2	Масований за кількістю та обсягом потік вхідних даних	Відсутність фільтрації та обмеження кількості вхідних запитів	Відмова в обслуго-вуванні	Порушення доступності системи	Висока	Середня	Високий	Обмеження кількості запитів
3	Експлоїти до СШІ та їх оточення	Компрометація СШІ або її оточення, збереження даних у вихідному вигляді	Вхідний витік	Втрата даних, порушення конфіденції ності	Висока	Низька	Серед-ній	Шифруван-ня даних
4	Дані, що подаються на вхід системи для аналізу	Вихідні обмеження СШІ (навчання ґрунтується на вивченні патернів), недостатній обсяг навчальних даних	Змагаль ні атаки	Втрата цілісності системи	Висока	Висока	Високий	Змагальні навчання, введення у дані змагальних прикладів, модифікація вхідних даних, захисна дистиляція, метод вилучення інваріантів, стискання функцій

Продовження табл. 2

1	2	3	4	5	6	7	8	9
5	Звичайні вхідні дані, які система навчена класифікувати неправильно	Використання сторонніх даних, ненадійні дані, відсутня фільтрація даних	Отруєння даних	Втрата цілісності системи	Висока	Висока	Високий	Користуватися даними з надійних джерел, валідувати та фільтрувати дані, захищати їх під час використання
6	Запити до системи, націлені на збирання даних, на яких вона навчалася	Принципи навчання СШІ, залежність від зовнішніх даних	Витік даних	Втрата даних, порушення конфіденційності	Висока	Висока	Високий	Обмеження кількості запитів

На підставі результатів аналізу атак за рівнем небезпеки для СШІ побудуємо матрицю критичності (кіберризиків) цих систем (табл. 3) та матрицю критичності після впровадження контрзаходів (табл. 4). Зеленим позначено низький рівень ризику (атака 1), жовтим – середній рівень ризику (атака 3), червоним – високий (атаки 2, 4, 5, 6).

Таблиця 3

Матриця критичності кіберризиків СШІ

	Тяжкість		
Ймовірність появи	Низька	Середня	Висока
Низька			
Середня	1		
Висока	3	2, 6	4, 5

Таблиця 4

Матриця критичності кіберризиків СШІ після упровадження контрзаходів

	Тяжкість		
Ймовірність появи	Низька	Середня	Висока
Низька		2, 6	
Середня	1		
Висока	3		4, 5

На підставі аналізу матриці критичності (табл. 4) атаки “Відмова в обслуговуванні” (2) та “Витік даних” (6) змінюють рівень ймовірності появи з високого на низький завдяки ефективним контрзаходам. Однак “Змагальні атаки” (4) та “Отруєння даних” (5) залишають системи в зоні високого ризику, бо наявні контрзаходи ніяк не толерують наслідки цих атак.

4. Загальні методи і рекомендації щодо забезпечення кібербезпеки ШІ, глобальна міждержавна взаємодія із питань безпечності ШІ

Останніми роками спостерігається зростання уваги на рівні урядів до ШІ та його безпеки [39, 40]. Фахівці країн, які розуміють виклики у сфері кібербезпеки СШІ, підкреслюють важливість прозорості, тестування та підзвітності алгоритмів та їх розробників [39]. Наприклад, у Сполучених Штатах Америки Комісія із національної безпеки штучного інтелекту (NSCAI) наголосила на важливості створення гарантоздатних (надійних і безпечних) СШІ, які можна перевіряти за допомогою суворої стандартизованої системи документації [41]. З цією метою комісія рекомендувала розробити стандарти для моделей ШІ, ураховуючи вимоги щодо того, які дані використовують моделі, які параметри та їхня вага для моделей, як вони навчаються і тестуються та які результати отримують. Це дасть змогу експертам виявляти вразливості технології ШІ, ризики потенційного маніпулювання із вхідними даними, а також інших неочікуваних результатів [39].

Регуляторні органи також відіграють важливу роль у цьому процесі. Вони можуть розробити механізми підзвітності та режими відповідальності для управління ШІ у разі виникнення кіберінцидентів. Це можуть бути базові вимоги до розробників СШІ щодо, наприклад, отримання сертифікатів, проходження аудитів та тестування з урахуванням специфічних характеристик штучного інтелекту. Розробники, які не дотримуються цих стандартів, створюючи СШІ, у випадку їх компрометації відповідатимуть за заподіяну шкоду [39].

Стосовно захисту СШІ доцільно сформулювати такі основні рекомендації:

- передусім потрібно впроваджувати безпечний життєвий цикл розроблення СШІ [6, 42, 43];
- “традиційні” типи вразливостей ПЗ можна толерувати, мінімізуючи використання стороннього коду чи ретельно перевіряючи захищеність коду, без якого неможливо обійтися у кожній конкретній ситуації;
- атакам на платформу можна протистояти, приділяючи більшу увагу тому, як обробляються вхідні дані, захищаючи платформу стандартними механізмами боротьби із DoS атаками та контролюючи оточення системи щодо витоку вхідних даних користувачів [8];
- атакам на алгоритм (змагальним атакам) можна протистояти завдяки навчанню, яке вводить у набір даних змагальні приклади [8, 44, 45], а також модифікації вхідних даних [46]. Також є специфічні методи боротьби із цими атаками, як наприклад, захисна дистиляція [47], метод вилучення інваріантів [48], стискання функцій [49] тощо;
- щоб протистояти атакам на дані, треба використовувати нескомпрометовані набори даних з надійних джерел, валідувати та фільтрувати ці дані, а також захищати їх під час використання. Дані мають бути не персоналізовані, щоб уникнути проблем конфіденційності. Крім того, лінійна регресія може допомогти у боротьбі з атаками отруєння даних [50]. Існують і специфічніші методи для боротьби з цими атаками [26].

5. Висновки за результатами аналізу

На підставі аналізу стану речей щодо кібербезпеки СШІ виявлено, що вони вразливі як для класичних атак на програмне забезпечення, так і для специфічних векторів атак, притаманних тільки цим складним системам. Аналіз класичних вразливостей програмного забезпечення під час дослідження не здійснювався, оскільки цей напрям вже опрацьовували протягом багатьох років і він не є унікальним для СШІ.

Визначено, що специфічні вектори атак на СШІ складаються із трьох основних груп: “Атаки на платформу”, “Атаки на алгоритм” та “Атаки на дані”. “Атаки на платформу” за сутністю дуже близькі до класичних атак на програмне забезпечення. Модифікація даних, відмова в обслуговуванні, вхідний витік – це напрями, вже знайомі для спеціалістів з інформаційної безпеки. Існує багато методів боротьби з цими типами атак. Тому, враховуючи й те, що тяжкість їх наслідків, за нашими оцінками, на рівні, нижчому від середнього, цей напрям також не буде пріоритетним в подальших дослідженнях.

З іншого боку, “Атаки на алгоритм”, а саме “Змагальні атаки”, – це абсолютно новий тип атак, специфічний тільки для СШІ. Через високий рівень ймовірності появи ці атаки створюють велику загрозу і рівень тяжкості їхніх наслідків теж високий. Однак основні контрзаходи щодо цих атак суто експериментальні (емпіричні). До цих контрзаходів можна зарахувати змагальне навчання, модифікацію вхідних даних (очищення від стороннього шуму) тощо. До того ж існує багато специфічних методів боротьби з цими загрозами, проте кожен з них охоплює доволі вузький діапазон можливих атак. Це ускладнює розроблення надійного захисту СШІ від цього типу атак.

Другим специфічним типом атак на СШІ є “Атаки на дані”, які складаються із двох підтипів “Отруєння даних” та “Витік даних”. Для здійснення атак на “Витік даних” потрібно здійснити велику кількість запитів до системи, що доволі вільно контролюють і толерують, обмежуючи кількість цих запитів (як це роблять для захисту від атак “Відмови в обслуговуванні”). Тому цей напрям має середній рівень тяжкості наслідків і не є пріоритетним для дослідження. Однак високий рівень ймовірності появи типу атак “Отруєння даних”, що істотно збільшує їх наслідки для СШІ. Протистояння цій загрозі є рекомендаційним і полягає у використанні даних з надійних джерел, їх валідуванні та фільтрації, захисті під час навчання. На нашу думку, цих рекомендацій недостатньо, щоб створити СШІ, захищені від таких атак.

Важливим є висновок про потребу в безпечному життєвому циклі розроблення СШІ. Для того, щоб не гальмувати застосування ШІ, до некритичних систем можна застосувати менш жорсткі рекомендації та гнучкий життєвий цикл, а для критичних необхідна доволі сувора та стандартизована система. Підкреслимо важливість прозорості, тестування, підзвітності алгоритмів та людей, які їх розробляють і, за певних умов, надають послуги із використанням СШІ. У разі відмов у критичних СШІ розробники та операційники повинні відповідати за заподіяну шкоду. Однак ця рекомендація потребує детального відпрацювання, зважаючи на доволі відчутні наслідки.

Безперечно, СШІ створюють нові цінності для суспільства. Тому для подальшого розвитку вкрай важливо підтримувати безпечність технологій штучного інтелекту як на етапі розроблення, так і на етапі експлуатації СШІ.

6. Обґрунтування напрямів дослідження

Ураховуючи результати попереднього аналізу, виділимо напрями майбутніх досліджень.

1. “Змагальні атаки” на ШІ дуже небезпечні для цих систем, їх рівень тяжкості за наслідками високий. А відомі контрзаходи здебільшого експериментальні або охоплюють вузький діапазон цих атак. Тому виникає необхідність підвищити показник стійкості ШІ до атак цього типу.

2. Рівень тяжкості атак “Отруєння даних” також має високий і загрожує системам ШІ ризиками втрати працездатності. Більшість відомої інформації стосовно захисту від атак цього типу суто рекомендаційна. Отже, підвищення стійкості систем ШІ до атак цього типу також є перспективним напрямом досліджень.

3. Створення моделі безпечного життєвого циклу розроблення і розгортання систем ШІ. Це відповідає сталим тенденціям і рішенням у галузі функційної та кібербезпеки. Для СШІ розроблено моделі якості [51, 52], важливою складовою яких є безпекові характеристики. Тому доцільне поєднання таких моделей і моделей життєвого циклу безпеки для таких систем.

4. Метод IMECA та інші відомі методи напівформального аналізу безпеки можуть використовуватися для ризик-орієнтованого оцінювання і вибору раціональних контрзаходів за критерієм “прийнятний ризик – вартість”. Доцільно удосконалити й адаптувати ці методи із урахуванням специфічних особливостей СШІ.

Отже, метою досліджень повинно бути підвищення достовірності оцінювання та захищеності СШІ на підставі розроблення і упровадження нових моделей якості, життєвого циклу безпеки, а також методів аналізу і забезпечення кібербезпеки на різних етапах розроблення і використання ШІ як сервісу (Artificial Intelligence as a Service, AIaaS [53]).

Список літератури

1. Herping S. (2019). Securing Artificial Intelligence – Part I. https://www.stiftung-nv.de/sites/default/files/securing_artificial_intelligence.pdf

2. PwC: The macroeconomic impact of artificial intelligence (2018). https://www.pwc.co.uk/economic-services/assets/macroeconomic-impact-of-ai-technical-report-feb-18.pdf

3. Comiter M. (2019). Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It. *Belfer Center for Science and International Affairs, Harvard Kennedy School*. https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf

4. Povolny S. (2020). Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles. *McAfee Labs*. https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/

5. Lohn A. (2020). Hacking AI. *Center for Security and Emerging Technology*. https://doi.org/10.51593/2020CA006

6. Lohn A. (2021). Poison in the Well. *Center for Security and Emerging Technology*. https://doi.org/10.51593/2020CA013

7. Ruef M. (2020). Hacking Artificial Intelligence – Influencing and Cases of Manipulation. https://www.researchgate.net/publication/338764153_Hacking_Artificial_Intelligence_-_Influencing_and_Cases_of_Manipulation

8. Kim A. (2020). The Impact of Platform Vulnerabilities in AI Systems. *Massachusetts Institute of Technology*. https://dspace.mit.edu/bitstream/handle/1721.1/129159/1227275868-MIT.pdf

9. Hartmann K., Steup C. (2020). Hacking the AI – the Next Generation of Hijacked Systems. *In 12 International Conference on Cyber Conflict (CyCon)*. https://doi.org/10.23919/CyCon49761.2020.9131724

10. Bursztein E. (2018). Attacks against machine learning – an overview. *Personal Site and Blog featuresing blog posts publications and talks*. https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

11. Ansah H. (2021). Adversarial Attacks on Neural Networks: Exploring the Fast Gradient Sign Method. *Neptune blog*. https://neptune.ai/blog/adversarial-attacks-on-neural-networks-exploring-the-fast-gradient-sign-method

12. Griffin J. (2019). Researchers hack AI video analytics with color printout. https://www.securityinfowatch.com/video-surveillance/video-analytics/article/21080107/researchers-hack-ai-video-analytics-with-color-printout

13. Thys S., Ranst W. V., Goedemé T. (2019). Fooling automated surveillance cameras: adversarial patches to attack person detection. *arXiv preprint arXiv:1904.08653*. https://doi.org/10.48550/arXiv.1904.08653

14. Eykholt K., Evtimov I., Fernandes E., Li B., Rahmati A., Xiao C., Prakash A., Kohno T., Song D. (2018). Robust Physical-World Attacks on Deep Learning Models. *arXiv preprint arXiv:1707.08945*. https://doi.org/10.48550/arXiv.1707.08945

15. Eykholt K., Evtimov I., Fernandes E., Li B., Rahmati A., Tramer F., Prakash A., Kohno T., Song D. (2018). Physical Adversarial Examples for Object Detectors. *arXiv preprint arXiv:1807.07769*. https://doi.org/10.48550/arXiv.1807.07769

16. Su J., Vargas D. V., Sakurai K. (2019). Attacking convolutional neural network using differential evolution. *IPSJ Transactions on Computer Vision and Applications*. https://doi.org/10.1186/s41074-019-0053-3

17. Goodfellow I. J., Shlens J., Szegedy C. (2015). Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*. https://doi.org/10.48550/arXiv.1412.6572

18. Papernot N., McDaniel P., Goodfellow I. J. (2016). Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *arXiv preprint arXiv:1605.07277*. https://doi.org/10.48550/arXiv.1605.07277

19. Catak F. O., Yayilgan S. Y. (2021). Deep Neural Network based Malicious Network Activity Detection Under Adversarial Machine Learning Attacks. *In International Conference on Intelligent Technologies and Applications*, 280–291. https://doi.org/10.1007/978-3-030-71711-7_23

20. Volborth M. (2019). Detecting backdoor attacks on artificial neural networks. https://ece.duke.edu/about/news/detecting-backdoor-attacks-artificial-neural-networks

21. Vincent J. (2020). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

22. Ji Y., Liu Z., Hu X., Wang P., Zhang Y. (2019). Programmable Neural Network Trojan for Pre-Trained Feature Extractor. *arXiv preprint arXiv:1901.07766*. https://doi.org/10.48550/arXiv.1901.07766

23. Yang Z., Iyer N., Reimann J., Virani N. (2019). Design of intentional backdoors in sequential models. *arXiv preprint arXiv:1902.09972*. https://doi.org/10.48550/arXiv.1902.09972

24. Gu T., Dolan-Gavitt B., Garg S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*. https://doi.org/10.48550/arXiv.1708.06733

25. Biggio B., Nelson B., Laskov P. (2013). Poisoning Attacks against Support Vector Machines. *arXiv preprint arXiv:1206.6389*. https://doi.org/10.48550/arXiv.1206.6389

26. Jagielski M., Oprea A., Biggio B., Liu C., NitaRotaru C., Li B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. *In 2018 IEEE Symposium on Security and Privacy (SP)*, 19–35. https://doi.org/10.1109/SP.2018.00057

27. Xiao H., Biggio B., Brown G., Fumera G., Eckert C., Roli F. (2015). Is feature selection secure against training data poisoning? *In International Conference on Machine Learning*, 1689–1698. https://doi.org/10.48550/arXiv.1804.07933

28. Fredrikson M., Jha S., Ristenpart T. (2015). Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. *In CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–1333. https://doi.org/10.1145/2810103.2813677

29. Shokri R., Stronati M., Song C., Shmatikov V. (2017). Membership Inference Attacks against Machine Learning Models. *In the proceedings of the IEEE Symposium on Security and Privacy*. https://doi.org/10.48550/arXiv.1610.05820

30. Salem A., Zhang Y., Humbert M., Berrang P., Fritz M., Backes M. (2018). ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. *arXiv preprint arXiv:1806.01246*. https://doi.org/10.48550/arXiv.1806.01246

31. Rahman A., Rahman T., Lagani`ere R., Mohammed N., Wang Y. (2018). Membership Inference Attack against Differentially Private Deep Learning Model. https://www.tdp.cat/issues16/tdp.a289a17.pdf

32. Song L., Shokri R., Mittal P. (2019). Privacy Risks of Securing Machine Learning Models against Adversarial Examples. *arXiv preprint arXiv:1905.10291*. https://doi.org/10.48550/arXiv.1905.10291

33. Hayes J., Melis L., Danezis G., De Cristofaro E. (2018). LOGAN: Membership Inference Attacks Against Generative Models. *arXiv preprint arXiv:1705.07663*. https://doi.org/10.48550/arXiv.1705.07663

34. Singh P. (2022). Data Leakage in Machine Learning: How it can be detected and minimize the risk. https://towardsdatascience.com/data-leakage-in-machine-learning-how-it-can-be-detected-and-minimize-the-risk-8ef4e3a97562

35. Rakin A. S., He Z., Fan D. (2019). Bit-Flip Attack: Crushing Neural Network with Progressive Bit Search. *arXiv preprint arXiv:1903.12269*. https://doi.org/10.48550/arXiv.1903.12269

36. Tramèr F., Zhang F., Juels A., Reiter M. K., Ristenpart T. (2016). Stealing Machine Learning Models via Prediction APIs. *Proceedings of the 25th USENIX Security Symposium*. https://doi.org/10.48550/arXiv.1609.02943

37. Bhagoji A. N., Chakraborty S., Mittal P., Calo S. B. (2019). Analyzing Federated Learning through an Adversarial Lens. *In Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:634–643. http://proceedings.mlr.press/v97/bhagoji19a.html

38. Androulidakis I., Kharchenko V., Kovalenko A. (2016). IMECA-based Technique for Security Assessment of Private Communications: Technology and Training. https://doi.org/10.11610/isij.3505

39. Wolff J. (2020). How to improve cybersecurity for artificial intelligence. *The Brookings Institution*. https://www.brookings.edu/research/how-to-improve-cybersecurity-for-artificial-intelligence/

40. Newman J. C. (2019). Toward AI Security GLOBAL ASPIRATIONS FOR A MORE RESILIENT FUTURE. https://cltc.berkeley.edu/wp-content/uploads/2019/02/Toward_AI_Security.pdf

41. National Security Commission on Artificial Intelligence. First Quarter Recommendations. (2020). https://drive.google.com/file/d/1wkPh8Gb5drBrKBg6OhGu5oNaTEERbKss/view

42. Pupillo L., Fantin S., Ferreira A., Polito C. (2021). Artificial Intelligence and Cybersecurity. *CEPS Task Force Report*. https://www.ceps.eu/wp-content/uploads/2021/05/CEPS-TFR-Artificial-Intelligence-and-Cybersecurity.pdf

43. Neustadter D. (2020). Why AI Needs Security. *Synopsys Technical Bulletin*. https://www.synopsys.com/designware-ip/technical-bulletin/why-ai-needs-security-dwtb-q318.html

44. Tramèr F., Kurakin A., Papernot N., Goodfellow I., Boneh D., McDaniel P. (2020). Ensemble Adversarial Training: Attacks and Defenses. *arXiv preprint arXiv:1705.07204*. https://doi.org/10.48550/arXiv.1705.07204

45. Yuan X., He P., Zhu Q., Li X. (2018). Adversarial Examples: Attacks and Defenses for Deep Learning. *arXiv preprint arXiv:1712.07107*. https://doi.org/10.48550/arXiv.1712.07107

46. Dziugaite G. K., Ghahramani Z., Roy D. M. (2016). A study of the effect of JPG compression on adversarial images. *arXiv preprint arXiv:1608.00853*. https://doi.org/10.48550/arXiv.1608.00853

47. Papernot N., McDaniel P., Wu X., Jha S., Swami A. (2016). Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. *In 2016 IEEE Symposium on Security and Privacy (SP)*, 582-597. https://doi.org/10.1109/SP.2016.41

48. Ma S., Liu Y., Tao G., Lee W.C., Zhang X. (2019). NIC: Detecting Adversarial Samples with Neural Network Invariant Checking. *In NDSS*. https://www.ndss-symposium.org/ndss-paper/nic-detecting-adversarial-samples-with-neural-network-invariant-checking/

49. Xu W., Evans D., Qi Y. (2018). Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. *In Network and Distributed Systems Security Symposium (NDSS)*. https://doi.org/10.14722/ndss.2018.23198

50. Liu C., Li B., Vorobeychik Y., Oprea A. (2017). Robust linear regression against training data poisoning. *In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 91–102. https://doi.org/10.1145/3128572.3140447

51. Kharchenko V., Fesenko H., Illiashenko O. (2022). Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application. https://doi.org/10.3390/s22134865

52. Kharchenko V., Fesenko H., Illiashenko O. (2022). Basic model of non-functional characteristics for assessment of artificial intelligence quality. *Radioelectronic and computer systems*. https://doi.org/10.32620/reks.2022.2.11

53. Janbi N., Katib I., Albeshri A., Mehmood R. (2020). Distributed Artificial Intelligence-as-a-Service (DAIaaS) for Smarter IoE and 6G Environments. https://doi.org/10.3390/s20205796

References

1. Herping, S. (2019). Securing Artificial Intelligence – Part I. https://www.stiftung-nv.de/sites/default/files/securing_artificial_intelligence.pdf

2. PwC: The macroeconomic impact of artificial intelligence. (2018). https://www.pwc.co.uk/economic-services/assets/macroeconomic-impact-of-ai-technical-report-feb-18.pdf

3. Comiter, M. (2019). Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It. *Belfer Center for Science and International Affairs, Harvard Kennedy School*. https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf

4. Povolny, S. (2020). Model Hacking ADAS to Pave Safer Roads for Autonomous Vehicles. *McAfee Labs*. https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/

5. Lohn, A. (2020). Hacking AI. *Center for Security and Emerging Technology*. https://doi.org/10.51593/2020CA006

6. Lohn, A. (2021). Poison in the Well. *Center for Security and Emerging Technology*. https://doi.org/10.51593/2020CA013

7. Ruef, M. (2020). Hacking Artificial Intelligence – Influencing and Cases of Manipulation. https://www.researchgate.net/publication/338764153_Hacking_Artificial_Intelligence_-_Influencing_and_Cases_of_Manipulation

8. Kim, A. (2020). The Impact of Platform Vulnerabilities in AI Systems. *Massachusetts Institute of Technology*. https://dspace.mit.edu/bitstream/handle/1721.1/129159/1227275868-MIT.pdf

9. Hartmann, K., & Steup, C. (2020). Hacking the AI – the Next Generation of Hijacked Systems. *In 12 International Conference on Cyber Conflict (CyCon)*. https://doi.org/10.23919/CyCon49761.2020.9131724

10. Bursztein, E. (2018). Attacks against machine learning – an overview. *Personal Site and Blog featuresing blog posts publications and talks*. https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

11. Ansah, H. (2021). Adversarial Attacks on Neural Networks: Exploring the Fast Gradient Sign Method. *Neptune blog*. https://neptune.ai/blog/adversarial-attacks-on-neural-networks-exploring-the-fast-gradient-sign-method

12. Griffin, J. (2019). Researchers hack AI video analytics with color printout. https://www.securityinfowatch.com/video-surveillance/video-analytics/article/21080107/researchers-hack-ai-video-analytics-with-color-printout

13. Thys, S., Ranst, W.V., & Goedemé, T. (2019). Fooling automated surveillance cameras: adversarial patches to attack person detection. *arXiv preprint arXiv:1904.08653*. https://doi.org/10.48550/arXiv.1904.08653

14. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust Physical-World Attacks on Deep Learning Models. *arXiv preprint arXiv:1707.08945*. https://doi.org/10.48550/arXiv.1707.08945

15. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramer, F., Prakash, A., Kohno, T., & Song, D. (2018). Physical Adversarial Examples for Object Detectors. *arXiv preprint arXiv:1807.07769*. https://doi.org/10.48550/arXiv.1807.07769

16. Su, J., Vargas, D. V., & Sakurai, K. (2019). Attacking convolutional neural network using differential evolution. *IPSJ Transactions on Computer Vision and Applications*. https://doi.org/10.1186/s41074-019-0053-3

17. Goodfellow, I.J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*. https://doi.org/10.48550/arXiv.1412.6572

18. Papernot, N., McDaniel, P., & Goodfellow, I.J. (2016). Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *arXiv preprint arXiv:1605.07277*. https://doi.org/10.48550/arXiv.1605.07277

19. Catak, F.O., & Yayilgan, S.Y. (2021). Deep Neural Network based Malicious Network Activity Detection Under Adversarial Machine Learning Attacks. *In International Conference on Intelligent Technologies and Applications*, 280-291. https://doi.org/10.1007/978-3-030-71711-7_23

20. Volborth, M. (2019). Detecting backdoor attacks on artificial neural networks. https://ece.duke.edu/about/news/detecting-backdoor-attacks-artificial-neural-networks

21. Vincent, J. (2020). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

22. Ji, Y., Liu, Z., Hu, X., Wang, P., & Zhang, Y. (2019). Programmable Neural Network Trojan for Pre-Trained Feature Extractor. *arXiv preprint arXiv:1901.07766*. https://doi.org/10.48550/arXiv.1901.07766

23. Yang, Z., Iyer, N., Reimann, J., & Virani, N. (2019). Design of intentional backdoors in sequential models. *arXiv preprint arXiv:1902.09972*. https://doi.org/10.48550/arXiv.1902.09972

24. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*. https://doi.org/10.48550/arXiv.1708.06733

25. Biggio, B., Nelson, B., & Laskov, P. (2013). Poisoning Attacks against Support Vector Machines. *arXiv preprint arXiv:1206.6389*. https://doi.org/10.48550/arXiv.1206.6389

26. Jagielski, M., Oprea, A., Biggio, B., Liu, C., NitaRotaru, C., & Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. *In 2018 IEEE Symposium on Security and Privacy (SP)*, 19–35. https://doi.org/10.1109/SP.2018.00057

27. Xiao, H., Biggio, B., Brown, G., Fumera, G., Eckert, C., & Roli, F. (2015). Is feature selection secure against training data poisoning? *In International Conference on Machine Learning*, 1689–1698. https://doi.org/10.48550/arXiv.1804.07933

28. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. *In CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–1333. https://doi.org/10.1145/2810103.2813677

29. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks against Machine Learning Models. *In the proceedings of the IEEE Symposium on Security and Privacy*. https://doi.org/10.48550/arXiv.1610.05820

30. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., & Backes, M. (2018). ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. *arXiv preprint arXiv:1806.01246*. https://doi.org/10.48550/arXiv.1806.01246

31. Rahman, A., Rahman, T., Lagani`ere, R., Mohammed, N., & Wang, Y. (2018). Membership Inference Attack against Differentially Private Deep Learning Model. https://www.tdp.cat/issues16/tdp.a289a17.pdf

32. Song, L., Shokri, R., & Mittal, P. (2019). Privacy Risks of Securing Machine Learning Models against Adversarial Examples. *arXiv preprint arXiv:1905.10291*. https://doi.org/10.48550/arXiv.1905.10291

33. Hayes, J., Melis, L., Danezis, G., & De Cristofaro, E. (2018). LOGAN: Membership Inference Attacks Against Generative Models. *arXiv preprint arXiv:1705.07663*. https://doi.org/10.48550/arXiv.1705.07663

34. Singh, P. (2022). Data Leakage in Machine Learning: How it can be detected and minimize the risk. https://towardsdatascience.com/data-leakage-in-machine-learning-how-it-can-be-detected-and-minimize-the-risk-8ef4e3a97562

35. Rakin, A.S., He, Z., & Fan, D. (2019). Bit-Flip Attack: Crushing Neural Network with Progressive Bit Search. *arXiv preprint arXiv:1903.12269*. https://doi.org/10.48550/arXiv.1903.12269

36. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., & Ristenpart, T. (2016). Stealing Machine Learning Models via Prediction APIs. *Proceedings of the 25th USENIX Security Symposium*. https://doi.org/10.48550/arXiv.1609.02943

37. Bhagoji, A.N., Chakraborty, S., Mittal, P., & Calo, S.B. (2019). Analyzing Federated Learning through an Adversarial Lens. *In Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:634-643. http://proceedings.mlr.press/v97/bhagoji19a.html

38. Androulidakis, I., Kharchenko, V., & Kovalenko, A. (2016). IMECA-based Technique for Security Assessment of Private Communications: Technology and Training. https://doi.org/10.11610/isij.3505

39. Wolff, J. (2020). How to improve cybersecurity for artificial intelligence. *The Brookings Institution*. https://www.brookings.edu/research/how-to-improve-cybersecurity-for-artificial-intelligence/

40. Newman, J. C. (2019). Toward AI Security GLOBAL ASPIRATIONS FOR A MORE RESILIENT FUTURE. https://cltc.berkeley.edu/wp-content/uploads/2019/02/Toward_AI_Security.pdf

41. National Security Commission on Artificial Intelligence. First Quarter Recommendations (2020). https://drive.google.com/file/d/1wkPh8Gb5drBrKBg6OhGu5oNaTEERbKss/view

42. Pupillo, L., Fantin, S., Ferreira, A., & Polito, C. (2021). Artificial Intelligence and Cybersecurity. *CEPS Task Force Report*. https://www.ceps.eu/wp-content/uploads/2021/05/CEPS-TFR-Artificial-Intelligence-and-Cybersecurity.pdf

43. Neustadter, D. (2020). Why AI Needs Security. *Synopsys Technical Bulletin*. https://www.synopsys.com/designware-ip/technical-bulletin/why-ai-needs-security-dwtb-q318.html

44. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2020). Ensemble Adversarial Training: Attacks and Defenses. *arXiv preprint arXiv:1705.07204*. https://doi.org/10.48550/arXiv.1705.07204

45. Yuan, X., He, P., Zhu, Q., & Li, X. (2018). Adversarial Examples: Attacks and Defenses for Deep Learning. *arXiv preprint arXiv:1712.07107*. https://doi.org/10.48550/arXiv.1712.07107

46. Dziugaite, G. K., Ghahramani, Z., & Roy, D. M. (2016). A study of the effect of JPG compression on adversarial images. *arXiv preprint arXiv:1608.00853*. https://doi.org/10.48550/arXiv.1608.00853

47. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. *In 2016 IEEE Symposium on Security and Privacy (SP)*, 582–597. https://doi.org/10.1109/SP.2016.41

48. Ma, S., Liu, Y., Tao, G., Lee, W.C., & Zhang, X. (2019). NIC: Detecting Adversarial Samples with Neural Network Invariant Checking. *In NDSS*. https://www.ndss-symposium.org/ndss-paper/nic-detecting-adversarial-samples-with-neural-network-invariant-checking/

49. Xu, W., Evans, D., & Qi, Y. (2018). Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. *In Network and Distributed Systems Security Symposium (NDSS)*. https://doi.org/10.14722/ndss.2018.23198

50. Liu, C., Li, B., Vorobeychik, Y., & Oprea, A. (2017). Robust linear regression against training data poisoning. *In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 91–102. https://doi.org/10.1145/3128572.3140447

51. Kharchenko, V., Fesenko, H., & Illiashenko, O. (2022). Quality Models for Artificial Intelligence Systems: Characteristic-Based Approach, Development and Application. https://doi.org/10.3390/s22134865

52. Kharchenko, V., Fesenko, H., & Illiashenko, O. (2022). Basic model of non-functional characteristics for assessment of artificial intelligence quality. *Radioelectronic and computer systems*. https://doi.org/10.32620/reks.2022.2.11

53. Janbi, N., Katib, I., Albeshri, A., & Mehmood, R. (2020). Distributed Artificial Intelligence-as-a-Service (DAIaaS) for Smarter IoE and 6G Environments. https://doi.org/10.3390/s20205796

ENSURANCE OF ARTIFICIAL INTELLIGENCE SYSTEMS CYBER SECURITY: ANALYSIS OF VULNERABILITIES, ATTACKS AND COUNTERMEASURES

Oleksii Neretin[1], Vyacheslav Kharchenko[2]

[1, 2] National aerospace university "KhAI",
[1] oleksii.neretin@gmail.com, ORCID 0000-0003-2114-6714,
[2] v.kharchenko@csn.khai.edu, ORCID 0000-0001-5352-077X

In recent years, many companies have begun to integrate artificial intelligence systems (AIS) into their infrastructures. AIS is used in sensitive areas of society, such as the judicial system, critical

infrastructure, video surveillance, and others. This determines the need for a reliable assessment and guaranteed provision of cyber security of AIS. The study analyzed the state of affairs regarding the cyber security of these systems. Possible types of attacks are classified and the main ones are considered in detail. Threats and attacks were analyzed by level of severity and security risks were assessed using the IMECA method. "Adversarial attacks" and "Data poisoning" attacks are found to have the highest risks of danger, but the countermeasures are not at the appropriate level. It was concluded that there is a need for formalization and standardization of the life cycle of the development and use of secure AIS. The directions of further research regarding the need to develop methods for evaluating and ensuring cyber security of the AIS are substantiated, including for systems that provide AI as a service.

Key words: artificial intelligence; cyber security; adversarial attacks; poisoning and data leakage; trojan attacks; model attacks; model theft and poisoning; countermeasures.