Прим. № ___

ЗАТВЕРДЖЕНО

Наказ Адміністрації Державної служби спеціального зв’язку та захисту інформації України

___ _____ 20___ року № ___

Рекомендації
з кіберзахисту інформаційно-комунікаційних систем, які використовують технології штучного інтелекту

І. Загальні положення

1. Рекомендації з кіберзахисту інформаційно-комунікаційних систем, які використовують технології штучного інтелекту (далі – Рекомендації) розроблено на виконання пункту 2 плану заходів з реалізації Концепції розвитку штучного інтелекту в Україні на 2025 – 2026 роки, затвердженого розпорядженням Кабінету Міністрів України від 09 травня 2025 року № 457-р.

2. Рекомендації можуть використовуватися власниками та/або розпорядниками інформаційних, електронних комунікаційних, інформаційно-комунікаційних і технологічних систем (далі − ІКС), які використовують технології штучного інтелекту (далі – ШІ), під час розробки плану кіберзахисту, що включає описи поточного та/або цільового стану кіберзахисту.

3. У цих Рекомендаціях терміни вживаються в такому значенні:

дистиляція − метод стиснення моделі ШІ, який використовується для передання знань від більш точної, складнішої моделі ШІ до простішої моделі ШІ, при якому простіша модель ШІ залишається набагато меншою, швидшою та менш вимогливою до обчислювальних ресурсів, зберігаючи при цьому високий рівень точності;

диференціальна конфіденційність – підхід щодо анонімності та конфіденційності даних, який використовується для гарантування того, що аналіз великих наборів даних моделей ШІ або навчання моделей ШІ не розкриває окремо вибрані персональні дані в цьому наборі даних;

метод ансамблевого навчання – підхід машинного навчання, що полягає у використанні відповідей (рішень) від кількох незалежно навчених моделей ШІ для формування єдиного результату, що використовуватиметься як еталонний і спрямований на підвищення загальної точності, надійності та стійкості моделі ШІ шляхом зменшення ризику перенавчання та мінімізації помилок;

метод суперечливого навчання – підхід машинного навчання, що базується на використанні суперечливих даних при навчанні моделі ШІ, який дозволяє відрізняти ймовірні зловмисні та/або аномальні вхідні дані та відповідно на них

реагувати для підвищення стійкості та надійності моделі ШІ згідно зі встановленими до неї вимогами;

метод федеративного навчання – підхід машинного навчання, який дозволяє навчати модель ШІ на основі обсягу даних, розподілених між децентралізованими пристроями, при цьому дані ніколи не покидають ці пристрої;

перехресна валідація − статистичний метод оцінки моделі машинного навчання, який використовується для надійної оцінки її узагальнюючої здатності та запобігання перенавчанню і полягає у багаторазовому ітераційному поділі вхідного набору даних: для навчання моделі машинного навчання та для набору даних для оцінки продуктивності моделі машинного навчання на даних, що не використовувалися під час її навчання;

суперечливі дані – спеціально створена вхідна інформація для моделі ШІ, що викликає некоректну (помилкову) класифікацію, прогнозування та/або генерування відповідей моделлю ШІ, порушуючи тим самим функціональну цілісність вхідної інформації та кінцевих відповідей моделі ШІ.

Інші терміни у цих Рекомендаціях вживаються у значеннях, наведених у Законах України «Про основні засади кібербезпеки України», «Про захист інформації в інформаційно-комунікаційних системах», «Про електронні комунікації», Положенні про організаційно-технічну модель кіберзахисту, затвердженому постановою Кабінету Міністрів України від 29 грудня 2021 року № 1426, Національному плані реагування на кіберінциденти, кібератаки та кіберзагрози, затвердженому постановою Кабінету Міністрів України від 26 листопада 2025 року № 1533, Мінімальних вимогах до захисту інформаційних, електронних комунікаційних, інформаційно-комунікаційних та технологічних систем, затверджених постановою Кабінету Міністрів України від 29 березня 2006 року № 373.

4. Рекомендації не є нормативно-правовим актом, мають інформаційний та рекомендаційний характер, не встановлюють правових норм і є добровільними для використання.

Вони пропонують таксономію характерних (специфічних) для ІКС із ШІ кіберзагроз і заходів з кіберзахисту, які забезпечують зменшення (усунення) негативного впливу від реалізації таких кіберзагроз, яка не є вичерпною і може бути адаптована залежно від потреб і ресурсів суб’єкта забезпечення кібербезпеки та специфіки конкретної ІКС із ШІ.

Використання Рекомендацій є рекомендованим при ідентифікації та оцінюванні ризиків кібербезпеки ІКС із ШІ.

ІІ. Упровадження ІКС із ШІ

1. Для забезпечення якісного впровадження заходів з кіберзахисту в ІКС із ШІ проводяться підготовчі дії, визначені в пунктах 2 – 8 цього розділу.

2. Спершу визначаються цілі використання технологій штучного інтелекту в інформаційно-комунікаційних системах суб’єкта забезпечення кібербезпеки. Це передбачає відповіді на такі запитання:

які завдання повинні виконувати ІКС із ШІ?;

які завдання в ІКС із ШІ мають виконуватися за допомогою технологій ШІ?;

у чому різниця виконання завдань ІКС із ШІ, якщо б такі завдання виконувалися без ШІ (наприклад, часова різниця, різниця у використанні ресурсів ІКС, різниця у залученості персоналу для виконання завдань), оскільки технології ШІ мають бути каталізатором виконання завдань, а не інгібітором?;

чи можливе виконання зазначених завдань без використання технологій ШІ?;

які саме технології ШІ будуть використовуватися в ІКС із ШІ, їх можливості, особливості використання та ризики їх використання?;

Відповіді на зазначені вище запитання мають бути точними та вимірними для оцінки ефективності ІКС із ШІ та їх використання персоналом.

Результати аналізу цілей використання технологій ШІ в ІКС із ШІ можуть бути задокументовані у вигляді окремого звіту та/або розділу політики використання ШІ.

3. Власник та/або розпорядник ІКС із ШІ може залучати кваліфікованих фахівців (команди фахівців) з розробки та/або впровадження технологій ШІ в ІКС із ШІ. Вони можуть виконувати такі функції:

виконання ролі технічного адміністратора ІКС із ШІ;

регулярне тестування ІКС із ШІ та окремо використаних технологій ШІ (наприклад, використання тестових наборів даних для оцінки точності та надійності моделей ШІ, моделювання деструктивних дій та потенційно шкідливих сценаріїв впливу на модель ШІ з метою виявлення вразливостей/загроз, забезпечення перевірки моделей на наявність будь-яких упереджень або помилок);

періодичне оновлення та перенавчання моделей ШІ новими даними для запобігання погіршення якості (деградації) виконання завдань ІКС із ШІ;

консультативна допомога користувачам щодо використання технологій ШІ в ІКС із ШІ.

Зазначеним користувачам рекомендовано постійно підвищувати кваліфікацію, зокрема щодо особливостей використання сучасних технологій ШІ та нових кіберзагроз, пов’язаних із ними.

4. Для ефективної роботи ІКС із ШІ критично важливим є забезпечення якості та релевантності даних. Якість даних можна визначити як ступінь, до якого дані є точними, повними, послідовними, своєчасними та доступними для використання, тоді як релевантність даних означає, наскільки дані відповідають конкретному завданню, яке має вирішити ІКС із ШІ.

5. Для забезпечення якості даних можна впровадити деякі інструменти та певні метрики для ІКС із ШІ, зокрема:

кількість помилок або відношення правильних відповідей до неправильних (ступінь відповідності даних реальному стану (факту) або достовірному джерелу);

кількість пропущених відповідей (чи всі необхідні значення наявні в наборі даних);

відсоток відповідей, що не відповідають логічним правилам або умовам запиту, відносно різних місць зберігання даних (наприклад, у різних полях одного запису);

середній час між фактичною подією та її реєстрацією в ІКС із ШІ та/або частота оновлення даних (переважно для часових рядів);

відсоток відповідей, що дублюються або копіюються.

6. Для забезпечення релевантності даних можна впровадити певні метрики для ІКС із ШІ, зокрема:

обсяг наданих відповідей (порівняння кількості доступних відповідей ШІ з мінімально необхідним обсягом даних, визначеним для конкретної ІКС із ШІ);

відповідність розподілу (її допустимий показник) (порівняння навчальних наборів даних з отриманими відповідями ШІ);

час актуальності даних (чи є необхідні для ІКС із ШІ дані актуальними з погляду часу або ж визначення періоду такої актуальності);

повнота даних (відповідність кількості ключових ознак (змінних) наданих відповідей кількості необхідних ключових ознак відповідей, необхідних для виконання завдань ІКС із ШІ).

Можуть бути розроблені та впроваджені чіткі політики та процедури щодо збору даних, їх зберігання, модифікації та подальшої обробки, а також рекомендовано включати регулярний контроль (аудит), визначення ролей і відповідальності за управління даними в процесі експлуатації ІКС із ШІ.

7. У процесі забезпечення якості та релевантності даних власник та/або розпорядник ІКС із ШІ враховує ризики виникнення упередженості у даних та моделях ШІ включно з упередженістю, що може:

бути наслідком навмисного або ненавмисного «отруєння» даних;

спричиняти некоректні рішення моделей та технологій ШІ, які впливають на безпеку ІКС із ШІ;

знижувати стійкість моделі ШІ та ІКС із ШІ відповідно до деструктивного впливу та загроз;

створювати передумови для компрометації ІКС із ШІ через спрямовані маніпуляції набором даних;

бути наслідком використання іноземних продуктів, які не є адаптованими до українських реалій.

Власник та/або розпорядник ІКС із ШІ забезпечує:

документування можливих джерел упередженості у даних моделі ШІ (DatasetBias, SamplingBias, HistoricalBias);

проведення регулярного тестування моделей ШІ на ознаки упередженості;

здійснення перехресної валідації моделей ШІ з використанням незалежних наборів даних;

виявлення підозрілих відповідей моделі, які можуть свідчити про упередженість даних;

коригування навчальних наборів даних або повторне навчання моделі ШІ у разі виявлення значущої упередженості.

8. Власник та/або розпорядник ІКС із ШІ може розробляти політику використання ШІ, що має базуватися на етичних принципах, дотриманні законодавчих та регуляторних вимог.

Основними компонентами такої політики можуть бути:

управління даними та конфіденційність;

етична розробка та зменшення упереджень;
прозорість та документування;
ролі, відповідальність і нагляд;
моніторинг і контроль.

9. Власникам та/або розпорядникам ІКС із ШІ рекомендовано проводити періодичну комплексну оцінку ризиків кібербезпеки, пов’язаних з технологіями ШІ, в тому числі ризиків забезпечення безпеки ланцюга постачання технологій ШІ. За результатами обробки ризиків визначається необхідність впровадження заходів, що дозволяють зменшити (усунути) негативний вплив реалізації кіберзагроз. Процес оцінювання ризиків кібербезпеки, пов’язаних з використанням технологій ШІ в ІКС із ШІ, рекомендовано інтегрувати із загальною системою управління ризиками суб’єкта забезпечення кібербезпеки.

Упровадження процесу управління ризиками для ІКС із ШІ рекомендовано здійснювати відповідно до спеціалізованих для ШІ міжнародних стандартів, наприклад:

ISO/IEC 23894:2023 Guidanceon Risk Management for Artificial Intelligence;
ISO/IEC 42001:2023 Artificial Intelligence Management System (AIMS);
NIST AI 100-1 Artificial Intelligence Risk Management Framework (v.1.0);
NIST SP 800-218A: Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile.

III. Таксономія характерних (специфічних) для ІКС із ШІ кіберзагроз і заходів з кіберзахисту

1. Атаки на ланцюги постачання технологій ШІ − загрози, спрямовані на компрометацію апаратного та програмного забезпечення, що використовується для розробки, розгортання, тестування та роботи моделей ШІ.

Сценарії реалізації загрози атак на ланцюги постачання технологій ШІ:

експлуатація вразливостей програмних компонентів (використання вразливостей у вихідному коді програм, неактуальних (неоновлених) програмних бібліотек або попередньо скомпрометованих моделей ШІ);

атака на інтерфейс прикладного програмування (API) (маніпулювання запитами до API або використання слабких механізмів автентифікації чи викрадених облікових даних для отримання доступу до API, яка взаємодіє з моделлю ШІ або ІКС із ШІ);

фізична компрометація (компрометація апаратного забезпечення, що використовується для роботи моделей ШІ та ІКС із ШІ, наприклад, шляхом впровадження шкідливих компонентів у процесі виробництва або під час несанкціонованого доступу).

2. Заходи з кіберзахисту щодо загрози атак на ланцюги постачання технологій ШІ:

1) моніторинг постачальників:

перевірка проведення заходів з кіберзахисту щодо власних продуктів постачальників та партнерів;

перевірка проведення заходів з кіберзахисту щодо наданих продуктів

постачальників та партнерів, які є компонентами ІКС із ШІ або забезпечують її роботу;

моніторинг всіх компонентів ланцюга постачання, у тому числі залежності програмного забезпечення, його своєчасного та цілісного оновлення;

2) забезпечення цілісності та валідації даних:

розроблення та впровадження політики управління даними з ретельними протоколами перевірки для підтвердження якості та релевантності даних перед їх використання як навчального набору даних для моделі ШІ;

використання методів виявлення аномалій та винятків для ідентифікації та видалення підозрілих або помилкових даних;

проведення регулярних аудитів та очищення наборів даних для запобігання впливу шкідливих даних на навчання моделі ШІ.

3. «Отруєння» даних − навмисне внесення помилкових або спотворених даних до навчальної вибірки моделі ШІ з метою погіршення продуктивності моделі, формування некоректних чи упереджених результатів або створення прихованого доступу.

Сценаріями реалізації загрози «отруєння» даних є:

інфільтрація даних (отримання несанкціонованого доступу до навчальних наборів даних, використовуючи вразливості в процесах збору даних, реалізується через компрометацію сторонніх постачальників даних або використання доступу інсайдерів);

створення шкідливих даних (розроблення помилкових або спотворених зразків даних, які можуть містити нетипові мітки (позначки) або атрибути прихованого доступу);

введення шкідливих даних у навчальний набір даних для моделі ШІ: шкідливі дані вводяться під час початкового навчання моделі ШІ або постійного оновлення даних в ІКС із ШІ (за умови підтримки безперервного навчання з моделлю ШІ);

пошкодження моделі ШІ: навчання моделі ШІ на шкідливих даних, що призводить до упередженої та/або неправильної поведінки при використанні в ІКС із ШІ;

експлуатація вже скомпрометованої ІКС із ШІ: обхід засобів контролю безпеки ІКС із ШІ, створення шкідливих результатів або зниження надійності, ефективності використання ІКС із ШІ.

4. Заходи з кіберзахисту щодо загрози «отруєння» даних:

1) забезпечення цілісності та валідації даних:

розроблення та впровадження політики управління даними з ретельними протоколами перевірки для підтвердження якості та релевантності даних перед їх використання як навчального набору даних для моделі ШІ;

використання методів виявлення аномалій і винятків для ідентифікації та видалення підозрілих або помилкових даних;

проведення регулярних аудитів та очищення наборів даних для запобігання впливу шкідливих даних на навчання моделі ШІ;

2) моніторинг вхідних даних і поведінки моделі ШІ:

забезпечення контролю джерел даних та ефективності моделі ШІ на предмет підозрілих закономірностей або несподіваної поведінки, які можуть свідчити про спроби «отруєння» даних;

упровадження метрик та інструментів виявлення відхилень моделі для виявлення змін у продуктивності моделі ШІ, які можуть бути викликані отруєними даними;

3) упровадження надійних методів навчання моделей ШІ:

використання методу суперечливого навчання та методу федеративного навчання;

використання механізмів відхилення винятків для фільтрації потенційно шкідливих даних під час навчання;

упровадження процедури періодичного перенавчання моделі за допомогою перевірених наборів даних;

4) забезпечення контролю доступу та захисту даних:

запровадження обмеження доступу до навчальних наборів даних моделі ШІ за допомогою контролю доступу на основі ролей та багатофакторної автентифікації;

забезпечення шифрування навчальних даних під час зберігання та їх передачі;

забезпечення надійного та захищеного каналу передачі даних для запобігання введенню шкідливих даних із зовнішніх або внутрішніх джерел;

5) упровадження інструментів безпеки ШІ:

упровадження рішень щодо управління безпекою ШІ, які забезпечують комплексну видимість, аналіз шляхів атак і виявлення неправильних налаштувань в ІКС із ШІ.

5. «Отруєння» моделей ШІ − маніпуляція параметрами або архітектурою моделі ШІ під час навчання для вбудовування шкідливої поведінки.

Сценарій реалізації загрози «отруєння» моделей ШІ:

доступ до процесу навчання моделей ШІ: несанкціоноване отримання доступу до навчання моделей ШІ, що може відбуватися в середовищах спільного навчання, таких як федеративне навчання, або шляхом непрямого маніпулювання даними навчання;

маніпулювання навчанням моделей ШІ: несанкціоноване внесення змін у параметри навчання моделі, наприклад шкідливі оновлення;

експлуатація отруєних моделей ШІ: використання скомпрометованої ІКС ШІ для несанкціонованого доступу до даних, створення шкідливих результатів або зниження надійності та операційної ефективності системи.

6. Заходи з кіберзахисту щодо загрози «отруєння» моделей ШІ:

1) забезпечення якості та релевантності даних:

забезпечення перевірки оновлень навчальних наборів;

використання автоматизованих інструментів і програмних рішень для

забезпечення якості та релевантності даних;

використання статистичних методів виявлення аномалій для позначення незвичайних шаблонів даних, які можуть вказувати на спроби «отруєння» моделей ШІ;

2) упровадження інструментів виявлення аномалій:

упровадження інструментів моніторингу та аналізу поведінки, відповідей та процесів навчання моделі ШІ для виявлення відхилень;

упровадження системи оповіщення, щоб ініціювати реагування у разі виявлення аномальних відповідей моделі ШІ або зниження ефективності її функціонування;

3) забезпечення контролю доступу та захисту даних:

запровадження обмеження доступу до даних, середовищ і параметрів навчання моделей за допомогою рольового контролю доступу, принципи мінімальних привілеїв та багатофакторної автентифікації;

4) упровадження надійних методів навчання моделей ШІ:

використання методу суперечливого навчання та методу федеративного навчання;

5) моніторинг та оцінка моделей ШІ:

забезпечення постійного моніторингу моделей ШІ для виявлення ознак зниження ефективності або підозрілої поведінки;

використання перехресної валідації та різноманітних наборів даних для оцінки надійності моделі ШІ;

забезпечення регулярного перенавчання моделі за допомогою перевірених даних;

6) забезпечення підвищення обізнаності:

забезпечення навчання кваліфікованих фахівців (команди фахівців) з розробки та/або впровадження технологій ШІ в ІКС із ШІ щодо ризиків, пов'язаних із загрозою «отруєнням» моделей ШІ;

розроблення планів реагування на інциденти, що стосуються підозрілих випадків «отруєння» моделі, для забезпечення своєчасного виявлення та мінімізації негативних наслідків;

7) упровадження інструментів безпеки ШІ:

упровадження рішень щодо управління безпекою ШІ, які забезпечують комплексну видимість, аналіз шляхів атак і виявлення неправильних налаштувань в ІКС із ШІ.

7. Змагальні атаки − створення спеціальних вхідних даних, які змушують модель ШІ робити помилкові класифікації з високим рівнем впевненості.

Сценарій реалізації загрози змагальних атак:

експлуатація − створення суперечливих вхідних даних із застосуванням спеціальних алгоритмів, що використовують вразливості в межах прийняття рішень моделлю, заплутуючи навчені представлення.

8. Заходи з кіберзахисту щодо загрози змагальних атак:

1) упровадження надійних методів навчання моделей ШІ:

використання методу суперечливого навчання та методу федеративного навчання;

2) забезпечення попередньої обробки вхідних даних:

упровадження алгоритмів попередньої обробки вхідних даних для зменшення або приховування суперечливих даних перед використання моделлю ШІ;

3) упровадження надійної архітектури:

використання різних моделей ШІ або архітектурно надійних моделей, які менш вразливі до ворожих вхідних даних. Ансамблі можуть послабити вплив ворожих прикладів, що націлені на одну модель;

забезпечення навчання вторинної моделі імітувати поведінку оригіналу з пом'якшеними результатами, зменшуючи чутливість до невеликих збурень вхідних даних. Ця техніка може підвищити стійкість до суперечливих вхідних даних шляхом згладжування меж прийняття рішень у моделі ШІ;

4) моніторинг і виявлення:

упровадження системи моніторингу, яка здатна виявляти аномальні вхідні дані;

5) постійна оцінка безпеки:

регулярне тестування моделі на можливі змагальні атаки та оновлення засобів захисту.

9. Атака типу «промпт-ін'єкція» у запити моделі ШІ − введення шкідливих або маніпулятивних запитів та/або інструкцій (промптів) до генеративних моделей ШІ з метою обходу механізмів захисту, несанкціонованого отримання інформації та/або несанкціонованого доступу.

Сценарій реалізації загрози атаки типу «промпт-ін'єкція» у запити моделі ШІ:

експлуатація: маніпулювання «промпт-ін'єкціями» для зміщення акценту згенерованого контексту, що призводить до некоректних відповідей або витоку даних, або спонукання системи до виконання операцій, не передбачених сценарієм роботи моделі ШІ (наприклад, отримання доступу до внутрішніх даних чи виконання змін).

10. Заходи з кіберзахисту щодо загрози атаки типу «промпт-ін'єкція» у запити моделі ШІ:

1) фільтрація вхідних запитів моделі ШІ:

розроблення жорстких правил синтаксичної та семантичної перевірки запитів моделі ШІ;

використання спеціалізованих бібліотек для фільтрації «промпт-ін'єкцій» (наприклад, Open AI Moderation API, Lang Chain Guardrails тощо);

упровадження регулярних виразів, списку дозволених команд та обмежень

для довжини та складності запитів моделі ШІ;

використання підходів машинного навчання для розпізнавання аномалій у вхідних даних;

2) упровадження обмежень контексту:

обмеження обсягу і типу інформації, доступної для моделі ШІ для обробки кожного запиту;

використання ізольованого середовища для виконання запитів;

запровадження політик контролю контексту моделі ШІ для запобігання маніпуляціям;

3) контроль і моніторинг поведінки системи:

упровадження систем моніторингу на предмет непередбаченої поведінки чи неналежної відповіді моделі ШІ;

логування та аналіз запитів і відповідей для раннього виявлення потенційних загроз;

автоматичне блокування та/або відхилення підозрілих запитів;

4) забезпечення контролю доступу та захисту даних:

упровадження багатофакторної автентифікації для користувачів, що надсилають критичні або привілейовані запити;

розмежування прав доступу до функцій ІКС із ШІ відповідно до ролей користувачів;

5) упровадження інструментів безпеки ШІ:

інтеграція засобів автоматичного виявлення шаблонів атак в запитах.

11. Інверсія моделі ШІ − використання інформації про навчальні дані або внутрішню структуру моделі ШІ в ІКС із ШІ з подальшим їх використанням для отримання несанкціонованого доступу до інформації або компрометації самої ІКС із ШІ.

Сценарії реалізації загрози інверсії моделі ШІ:

збір інформації: отримання доступу до результатів запитів моделі ШІ або її правил, збирання вихідних значень та інших внутрішніх даних, що генеруються моделлю ШІ;

експлуатація: відновлення даних за допомогою алгоритмів інверсії, на основі яких була навчена модель ШІ (можлива побудова моделі інверсії, яка використовує прогнози або оцінки достовірності моделі ШІ як вхідні дані для реконструкції приблизних представлень вихідних даних навчання для інших моделей ШІ);

ескалація загрози: у разі успішної інверсії моделі ШІ можливе створення нових атак на цю модель ШІ, таких як подальші атаки типу «промпт-ін'єкція» в запити моделі ШІ.

12. Заходи з кіберзахисту щодо загрози інверсії моделі ШІ:

1) забезпечення контролю щодо доступу та захисту даних:

обмеження доступу до результатів (відповідей) моделі ШІ виключно авторизованими користувачами;

мінімізація розкриття детальних показників надійності, ймовірностей і додаткових внутрішніх даних;

упровадження багатофакторної автентифікації;

розмежування прав доступу до ІКС із ШІ відповідно до ролей користувачів;

упровадження політик автентифікації та авторизації для інтерфейсів, що обслуговують модель ШІ;

2) увпровадження диференціальної конфіденційності:

упровадження технік диференціальної конфіденційності під час навчання моделі для зменшення ризику витоку персональних даних (у тому числі спеціалізованих бібліотек типу Tensor Flow Privacy і подібних);

використання механізмів обмеження впливу окремих даних на модель ШІ (розділення навчальних даних моделі ШІ тощо);

3) моніторинг та аудит:

забезпечення постійного моніторингу аномальної активності запитів, що може свідчити про спроби зловмисної інверсії;

запровадження ведення журналів запитів і відповідей для подальшого аналізу інцидентів безпеки;

4) забезпечення перевірки та мінімізації даних:

попередня обробка та очищення навчальних даних моделі ШІ, маскування або видалення чутливих ідентифікаторів моделі ШІ;

використання даних, що дають змогу зменшити ризик витоку даних моделі ШІ;

мінімізація чутливої інформації, що міститься в навчальних наборах даних, до рекомендованого мінімуму;

5) тестування та оцінка стійкості:

забезпечення регулярного проведення оцінки вразливостей моделей ШІ до інверсії;

адаптація і вдосконалення алгоритмів захисту моделей ШІ відповідно до нових загроз;

6) упровадження інструментів безпеки ШІ:

застосування рішень для забезпечення конфіденційності даних моделей ШІ, які підтримують технології безпечного навчання;

використання додаткових простіших моделей ШІ, які генерують менше персональних даних.

13. Крадіжка моделі ШІ − несанкціоноване створення без прямого доступу до внутрішніх параметрів, архітектури або навчальних даних моделі ШІ її копії для подальшого несанкціонованого використання.

Сценарій реалізації загрози крадіжки моделі ШІ:

копіювання і навчання копії моделі ШІ на основі цієї моделі ШІ: надсилання великого обсягу спеціально створених вхідних даних до моделі ШІ та навчання копії моделі ШІ на основі вихідних даних.

14. Заходи з кіберзахисту щодо загрози крадіжки моделі ШІ:

1) забезпечення контролю доступу та захисту даних:

упровадження контролю доступу до правил і протоколів моделі ШІ та результатів роботи моделі ШІ;

обмеження швидкості та обсягів видачі вихідних даних за допомогою надійних методів автентифікації, наприклад API-ключів;

упровадження багатофакторної автентифікації;

2) проактивне виявлення та моніторинг:

упровадження моніторингу використання моделей ШІ на предмет підозрілої поведінки запитів;

упровадження системи виявлення аномалій та аналітики безпеки в режимі реального часу для спрацьовування сповіщень та автоматизованих відповідей;

використання пасток для викриття та аналізу дій порушників;

3) дистиляція та захист моделей:

використання дистиляції моделей ШІ для створення спрощених схожих версій моделей ШІ, які важче піддаються реверс-інжинірингу (зворотній розробці);

4) забезпечення підвищення обізнаності:

забезпечення навчання кваліфікованих фахівців (команди фахівців) з розробки та/або впровадження технологій ШІ в ІКС із ШІ щодо ризиків, пов'язаних із загрозою крадіжки моделі ШІ;

розроблення планів реагування на інциденти, що стосуються випадків крадіжки моделі ШІ, для забезпечення своєчасного виявлення та мінімізації негативних наслідків;

проведення регулярних аудитів безпеки та перевірки доступу користувачів.

Заступник директора департаменту –
начальник відділу Департаменту кіберзахисту
Адміністрації Держспецзв’язку Денис ЗУБКОВ