КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (30), 2025
Спеціальний випуск ISSN 2663 – 4023
Special issue
DOI 10.28925/2663-4023.2025.30.993
УДК 004.056.5

Крет Тарас Богданович
Старший викладач кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0000-0002-6333-3190
taras.b.kret@lpnu.ua

Марценюк Євгеній Віталійович
асистент кафедри захисту інформації
Національний Університет "Львівська Політехніка", Львів, Україна
ORCID: 0009-0009-2289-0968
yevhenii.v.martseniuk@lpnu.ua

ІНТЕГРОВАНИЙ ПІДХІД ДО МОДЕЛЮВАННЯ ЗАГРОЗ У СИСТЕМАХ ШТУЧНОГО ІНТЕЛЕКТУ

Анотація. У роботі обґрунтовано актуальність моделювання загроз для систем штучного інтелекту в умовах зростання автономності моделей та появи нових векторів атак. Показано, що традиційні методи не враховують специфіку штучного інтелекту, що створює потребу у комплексному підході, здатному охопити весь життєвий цикл системи. Методологічну основу інтегрованого підходу становить поєднання міжнародних стандартів і галузевих практик: ISO/IEC 42001:2023 забезпечує керованість і аудит, NIST AI RMF 1.0 задає процесний цикл Govern–Map–Measure–Manage, MITRE ATLAS наповнює моделі реалістичними сценаріями атак, CSA MAESTRO додає багатошарову архітектурну декомпозицію, а OWASP GenAI Security Project пропонує операційні артефакти та інструменти пріоритезації. Такий синтез дозволяє інтегрувати стратегічні політики, технічні таксономії та практичні плейбуки в єдиний керований процес. Запропонований підхід робить моделювання загроз безперервним і доказовим, забезпечуючи трасовність від загрози до контролю та метрик ефективності. Він враховує технічні й соціотехнічні ризики, включаючи вплив на користувачів і суспільство, та підтримує профільну адаптацію для різних типів систем – від LLM до агентних платформ. Інтеграція з CI/CD-процесами та автоматизація перевірок підвищують швидкість реагування та знижують витрати на безпеку. Наукова новизна полягає у формуванні цілісного бачення, що поєднує керованість, процесну дисципліну, архітектурний аналіз і операційні інструменти. Практична значущість полягає у можливості застосування підходу для розробки комплексних стратегій захисту, сумісних із міжнародними нормами та придатних до сертифікаційної перевірки. Інтегрований підхід створює основу для масштабного впровадження ШІ з доведеним рівнем безпеки та довіри під час моделювання загроз. Він не лише підвищує стійкість систем, а й формує стандартизований контур управління ризиками, що відповідає сучасним викликам кіберзахисту.

Ключові слова: штучний інтелект; моделювання загроз; ISO/IEC 42001; NIST AI RMF; MITRE ATLAS; CSA MAESTRO; OWASP GenAI; інтегрований підхід; управління ризиками; безпека ШІ.

ВСТУП

Розвиток систем штучного інтелект (ШІ) стрімко інтегрується у всі сфери діяльності – від фінансових сервісів і медицини до промислових систем. Використання ШІ забезпечує автоматизацію процесів, підвищення ефективності та прийняття рішень на основі великих обсягів даних тощо. Однак зростання складності та автономності таких

рішень призводить до появи нових векторів атак, які суттєво відрізняються від традиційних кіберзагроз [1].

Атаки на моделі ШІ, такі як отруєння навчальних даних (Poison Training Data), обхід (ухилення) моделі ШІ (Evade AI Model) та порушення цілісності моделі ШІ (Erode AI Model Integrity), можуть призвести до некоректних дій і компрометації систем. Це особливо критично для систем, що працюють у сферах з високими вимогами до безпеки, наприклад, транспорт, енергетика, охорона здоров'я. Крім того, ШІ часто інтегрують у багаторівневі системи, що включають сенсори, контролери, хмарні сервіси та інтерфейси користувача [2]. Така розподілена структура збільшує кількість точок атаки та ускладнює побудову комплексного захисту [3]. Традиційні методи моделювання загроз, розроблені для класичних ІТ-систем, не враховують специфіку атак на ШІ-компоненти. Це створює потребу у нових підходах, здатних охопити весь життєвий цикл ШІ – від збору даних до розгортання та експлуатації моделей. Відсутність єдиного стандарту безпеки для систем штучного інтелекту ускладнює вибір оптимального методу для моделювання загроз. Таким чином, питання порівняння методів моделювання загроз для ШІ-систем є актуальним як з наукової точки зору, так з практичної.

Існує кілька провідних методик, які можна застосовувати до моделювання загроз для систем штучного інтелекту, кожен з яких має власну концепцію та сферу застосування. В цій роботі розглянемо такі з них:

1. NIST AI RMF (Risk Management Framework) фокусується на управлінні ризиками та забезпеченні відповідності стандартам, пропонуючи вимоги для оцінки та мінімізації загроз [4].
2. CSA MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) орієнтований на створення онтології ризиків, охоплюючи як технічні, так і організаційні аспекти безпеки ШІ [5].
3. MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) методика, що базується на тактиках і техніках атак, виявлених у реальних сценаріях [6].
4. OWASP Top 10 for LLM Applications Security Project визначає найбільш критичні вразливості для мовних моделей, що є особливо актуальним у контексті генеративного ШІ [7].
5. ISO/IEC 42001 являється основою системи менеджменту штучного інтелекту [12].

Кожен із цих методів має різний рівень деталізації, покриття життєвого циклу ШІ та інтеграцію з контролями безпеки. Проте відсутність єдиного підходу ускладнює вибір оптимальної методики для моделювання загроз, особливо коли йдеться про багаторівневі системи з компонентами ШІ. Додатковим викликом є швидкий розвиток технологій, що призводить до появи нових типів атак, які не завжди враховані у поточних методиках. Це створює потребу у систематичному порівнянні існуючих методів, визначенні їх сильних і слабких сторін та можливості комбінування для досягнення максимальної ефективності. Аналіз дозволить сформувати рекомендації для практичного застосування та розробки комплексних стратегій захисту систем ШІ.

Аналіз останніх досліджень і публікацій. Моделюванням загроз служать для розуміння потенційних ризиків безпеки, щоб запобігти їх реалізації, дозволяє зрозуміти, які активи є цінними, які вектори атак можливі та які контрзаходи потрібні для зниження ризиків. Такий підхід підвищує стійкість системи та зменшує витрати на усунення вразливостей у майбутньому [8, 9]. Розвиток ШІ створює нові вектори атак, що робить безпеку систем ШІ критично важливою, а впровадження методів моделювання загроз стає стратегічною необхідністю для захисту даних в цих системах [10]. Завдяки цьому можна не лише знизити ймовірність успішної атаки, а й оптимізувати витрати на безпеку,

концентруючи ресурси на найбільш критичних точках ризику. Науково-практичні дослідження [4-7, 11, 12] демонструють зближення кількох взаємодоповнювальних підходів до моделювання загроз у ШІ, які описані в *NIST AI RMF 1.0, MITRE ATLAS, CSA MAESTRO, OWASP GenAI Security Project та ISO/IEC 42001*.

NIST AI RMF 1.0 задає організаційну структуру через етапи: *управління (govern)* або «забезпечення керованості» у контексті політик і процесів – *відображення (map)* або «визначення / опис / ідентифікація контексту» об'єктів, потоків даних і взаємозв'язків – *вимірювання (measure)* тобто застосування кількісних та якісних методів аналізу ризиків та моніторингу ефективності заходів контролю – *керування (manage)* тобто впровадження заходів контролю, моніторингу та вдосконалення. Дані функції дозволяють вбудувати моделювання загроз у життєвий цикл систем та відрізнити ШІ-ризики від традиційних програмних вразливостей [4]. NIST AI RMF 1.0 дозволяє інтегруватись з корпоративним ризик-менеджментом, у межах яких моделювання загроз розміщується на етапах Map / Measure із подальшим переходом до Manage у вигляді планів пом'якшень та контролів.

MITRE ATLAS надає знання про тактики, техніки та процедури на основі реальних атак на ШІ-системи, що є опорою для побудови правдоподібних сценаріїв противника під час моделювання загроз [6]. *MITRE* ATLAS підкреслюють практичну цінність підходу для red team та моделювання загроз, забезпечуючи еталонні сценарії та узагальнені заходи пом'якшення.

CSA MAESTRO пропонує метод моделювання загроз, що охоплює рівні від моделей і даних до інфраструктури, спостережності та екосистеми агентів. Такий підхід дає змогу відслідковувати міжрівневі залежності та формувати мінімальні контрольні набори для кожного шару, фіксуючи ланцюги впливу в багатокрокових сценаріях для агентних і оркестрованих архітектур [5].

OWASP GenAI Security Project акцентує увагу на операційних артефактах – таксономії ризиків, контрольних списках й практичних інструкціях, що прискорюють переходи від ідентифікації загроз до їх практичного стримування [7].

ISO/IEC 42001 виконує роль системи менеджменту ШІ (AIMS), формалізуючи політики, ролі, процеси, аудит і безперервне вдосконалення у відповідальному застосуванні ШІ [12]. Завдяки сумісності з іншими управлінськими стандартами ISO, ISO/IEC 42001 дозволяє інтеграцію NIST-процесів, OWASP-процедур та MAESTRO-шарів у єдиний цикл моделювання загроз, що піддається сертифікаційному нагляду.

Проблематика дослідження. Існуючі методики не дозволяють забезпечити цілісного підходу, оскільки фокусуються на різних цілях: ISO/IEC 42001 формалізує політики, аудит і безперервне вдосконалення, але не деталізує технічні сценарії; NIST AI RMF пропонує цикл Govern–Map–Measure–Manage для управління ризиками, проте залишається технологічно нейтральним; MITRE ATLAS концентрується на таксономії TTPs (Tactics, Techniques, and Procedures) для побудови сценаріїв атак, не охоплюючи організаційні аспекти; OWASP GenAI Security Project акцентує на критичних ризиках генеративного ШІ та надає контрольні списки, але не інтегрує їх у процесний підхід; CSA MAESTRO пропонує шаровий метод для складних багатоагентних архітектур, однак потребує узгодження з профілями ризику та стандартами керованості.

З огляду на це можна стверджувати, що моделювання загроз для систем штучного інтелекту залишаються фрагментованими, оскільки кожна з них орієнтована на окремі аспекти – від управлінських процесів і ризик-менеджменту до таксономій атак та архітектурного аналізу – без інтеграції в єдиний цілісний підхід

Мета роботи: проаналізувати існуючі методи моделювання загроз для систем ШІ, визначити сильні та слабкі сторони кожного методу та запропонувати єдиний спільний підхід до моделювання загроз для систем ШІ.

Для вирішення поставленої мети необхідно вирішити наступні завдання:

1. Провести порівняльний аналіз методів моделювання загроз.
2. Оцінити сильні і слабкі сторони кожного методу.
3. Визначити доцільність комбінованого застосування різних методів для формування єдиного підходу для моделювання загроз.

ОСНОВНА ЧАСТИНА

Методика моделювання загроз для систем ШІ суттєво відрізняється від класичних підходів, застосовуваних до традиційних ІТ-систем. У класичних методах, таких як STRIDE [13], основний акцент робиться на статичних компонентах, чітких межах довіри та передбачуваних сценаріях атак. Натомість системи ШІ характеризуються динамічністю, здатністю до навчання, автономністю та взаємодією з непередбачуваними середовищами, що створює нові вектори загроз, включно з отруєнням даних, маніпуляціями під час навчання та атакою через підказки.

Важливою відмінністю є необхідність багатошарового аналізу, що охоплює не лише технічні аспекти, а й когнітивні властивості моделей, їхню здатність до автономного прийняття рішень та взаємодії з іншими агентами. На відміну від традиційних систем, де загрози здебільшого зводяться до порушення конфіденційності, цілісності та доступності, у ШІ додаються ризики викривлення логіки, втрати узгодженості цілей та неконтрольованого поширення помилкових рішень [14]. Також виникає необхідність у безперервному моделюванні загроз протягом усього життєвого циклу ШІ, оскільки моделі змінюються під час експлуатації. Таким чином, методика для ШІ має бути гнучкою, інтегрованою та орієнтованою на активи, щоб враховувати як класичні, так і нові типи атак, властиві системам з навчанням і автономністю [15]. З огляду на це розглянемо вимоги та підходи, які існують для моделювання загроз у системах ШІ, для цього детальніше проаналізуємо міжнародні стандарти, галузеві вимоги та кращі світові практики.

ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system. Міжнародний стандарт, який встановлює вимоги до системи управління штучним інтелектом, орієнтованої на безпеку, надійність та відповідальність упродовж усього життєвого циклу ШІ. ISO/IEC 42001:2023 формалізує процеси оцінки та впливу ризиків, їхнього оброблення, що безпосередньо пов’язано з моделюванням загроз. Ключові пункти 6.1.2–6.1.4 визначають обов’язковість ідентифікації ризиків, аналізу ймовірності та наслідків, а також розробки планів реагування, що створює основу для побудови моделей загроз.

Стандарт вимагає врахування не лише технічних аспектів, а й соціотехнічних ризиків, включаючи потенційні негативні впливи на користувачів і суспільство. Розділи 8–10 закріплюють операційне впровадження, моніторинг ефективності та безперервне вдосконалення, що робить моделювання загроз процесом, а не одноразовою дією. ISO/IEC 42001:2023 не нав’язує конкретну методику (наприклад, STRIDE чи ATLAS), але вимагає, щоб обраний підхід був формалізований, повторюваний і забезпечував трасовність від загрози до контрольного заходу. В таблиці 1 наведено зв'язок вимог стандарту ISO/IEC 42001 з моделюванням загроз.

Таблиця 1

Зв'язок вимог стандарту ISO/IEC 42001 з моделюванням загроз

Розділ / Пункт	Пояснення вимоги	Зв’язок з моделюванням загроз
6.1.2 AI Risk Assessment	Ідентифікація ризиків, оцінка ймовірності та наслідків	Формує основу для побудови моделей загроз
6.1.3 AI Impact Assessment	Аналіз впливу на користувачів і суспільство	Додає соціотехнічний вимір до моделювання загроз
6.1.4 AI Risk Treatment	Планування оброблення ризиків (уникнення, зменшення, передача)	Забезпечує трасовність від загрози до контролю
8 Operation	Операційне впровадження заходів	Інтеграція результатів моделювання у процеси
9 Performance Evaluation	Моніторинг, аудит, менеджмент-рев’ю	Перевірка актуальності моделей загроз
10 Improvement	Безперервне вдосконалення (PDCA)	Робить моделювання загроз циклічним процесом

Особливу увагу приділено активо-орієнтованому аналізу, що охоплює дані, моделі, ланцюги постачання та інтеграції з зовнішніми сервісами. Стандарт також передбачає документування результатів моделювання загроз, включаючи карти активів, сценарії атак, критерії ризику та журнали рішень. Вимога внутрішнього аудиту та менеджмент-рев’ю гарантує перевірку актуальності моделей загроз і їх узгодження з бізнес-цілями. ISO/IEC 42001 інтегрує PDCA-цикл, що забезпечує адаптивність до нових технік атак, дрейфу моделей та змін у середовищі.

Таким чином, стандарт розглядає моделювання загроз як центральний елемент управління ризиками ШІ, який має бути безперервним, доказовим і інтегрованим у всі процеси організації. Фактично закладений у стандарті підхід зобов’язує організації демонструвати не разову матрицю загроз, а системний підхід, де моделювання підживлює дизайн, експлуатацію та вдосконалення. Це робить ISO/IEC 42001 ключовим орієнтиром для побудови комплексної методики моделювання загроз, сумісної з міжнародними нормами та придатної до сертифікаційної перевірки.

NIST AI RMF (Risk Management Framework). NIST AI RMF 1.0 описує підхід для ризик-менеджменту ШІ, у якому моделювання загроз прямо не означено, але фактично є послідовність Govern – Map – Measure – Manage як практичний механізм ідентифікації, аналізу та оброблення загроз на всіх стадіях життєвого циклу системи.

У*правління (govern)* або «забезпечення керованості» вимагає наявності політик, ролей, критеріїв ризику та культури відповідальності – саме тут встановлюються правила, за якими будуються моделі загроз (об’єкти захисту, межі довіри, прийнятний ризик, критерії ескалації) та їхня трасовність до управлінських рішень. *Відображення (map)* орієнтує на опис контексту застосування, системних меж, даних, зацікавлених сторін і впливів, що становлять основу для виведення правдоподібних векторів атак і зловживань (misuse) у ШІ-середовищах. *Керування (manage)* орієнтується на створення метрик, показників і процедур перевірки, які дозволяють розділяти загрози та ефективність контрзаходів, перетворюючи моделювання загроз на повторюваний, верифікований процес, а не разову дію.

NIST AI RMF 1.0 не зобов’язує до використання конкретної техніки моделювання загроз, але ставить вимогу, аби підхід був доказовим, повторюваним і відтворюваним, із чіткими критеріями прийнятності, припущеннями та джерелами невизначеності. NIST AI RMF 1.0 чітко розрізняє ризик-оцінку (ліквідність/наслідки/невизначеність) і ризик-менеджмент, що вимагає під час моделювання загроз не лише каталогізувати

загрози, а й зв'язувати їх із бізнес-цілями та обов'язками щодо зацікавлених сторін. В NIST AI RMF 1.0 наголошується на безперервності процесу: загрози для ШІ еволюціонують (дрейф даних і моделей, нові TTPs, зміни в оточенні), отже моделі загроз слід переглядати за результатами моніторингу та інцидентів. Важливо також, що документ визнає відмінності між ризиками ШІ та традиційного ПЗ, тому очікує покриття векторів, притаманних саме ШІ (маніпуляції навчальними даними, інверсія/екстракція моделей, ненадійні/непояснювані висновки, вплив на справедливість і приватність).

У інтеграції з інструментами NIST AI RMF 1.0 (зокрема практичним AI RMF Playbook) моделювання загроз стає частиною керованої документації: карти системи, матриці ризику, критерії, протоколи моніторингу й звіти оцінювання. NIST AI RMF 1.0 підкреслює вимогу до метрик і доказів: дані про ефективність контролів, частоту/тяжкість інцидентів і тренди мають підтверджувати, що змодельовані загрози дійсно обробляються. Окремий акцент зроблено на неоднозначності та невизначеності оцінок – пропонується системно документувати припущення і межі застосовності, щоб не переоцінювати точність моделей загроз. Критично, що NIST AI RMF 1.0 не пропонує «універсального каталогу» загроз, натомість задає можливість використання зовнішніх таксономій (напр., ATLAS) або внутрішніх бібліотек для збагачення сценаріїв. У підсумку, NIST AI RMF позиціонує моделювання загроз як центральну складову: від встановлення контексту й активів, через розділення ризиків за метриками, до впровадження і перевірки контролів. Підкреслюється, що організації мають адаптувати моделювання загроз до власних профілів ризику та рівнів зрілості, зберігаючи сумісність із загальною структурою.

MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems). База знань представлена у вигляді матриці тактик, технік і пом'якшення негативних наслідків, створена на основі реальних спостережень за атаками на ШІ-системи та демонстрацій red team, яка напряму орієнтує моделювання загроз на правдоподібні сценарії противника. На відміну від загальних підходів ризик-менеджменту, ATLAS надає операційну гранулярність: 15 тактик, близько 130 технік, 26 пом'якшень і десятки реалізованих проектів. Все це можна безпосередньо накладати на активи, потоки даних і інтерфейси конкретної ШІ-системи.

Матриця моделює «чому» (tactics) і «як» (techniques) поведінки зловмисника проти ШІ (наприклад, poisoning, model extraction, prompt injection) і пропонує пом'якшення негативних наслідків, які можна використовувати як контрольні гіпотези у дизайні та підтвердженні їх відповідності. ATLAS є доповненням до класичного MITRE ATT&CK®, однак фокусує увагу на унікальних вразливостях ШІ, роблячи її релевантною саме для моделювання загроз в ШІ-контексті (моделі, дані, оркестрація інструментів, LLM-вектори тощо). Практичність ATLAS підсилюється публічними проектами, що показують фактичні наслідки атак і дозволяють будувати сценарії з оцінкою впливу та доказами експлуатації вразливостей.

Для моделювання загроз ATLAS виконує роль таксономії TTPs і «каталогу» шаблонів, які можна імпортувати у власні моделі, забезпечивши узгоджену термінологію. З погляду вимог до моделювання, ATLAS фактично пропонує орієнтований на зловмисника підхід: ідентифікацію активів і довірчих меж, відображення релевантних тактик/технік, підбір пом'якшень і перевірку їхньої дієвості через контрольовані атаки. Матриця масштабується по доменах (LLM, CV, автономні агенти тощо) завдяки розширенню реалізованих проектів і технік, що дає змогу формувати профілі загроз для конкретних застосувань. Водночас ATLAS не є процесним стандартом, тому необхідне його поєднання з RMF/AIMS-підходами (наприклад, NIST/ISO 42001) для формалізації

ролей, метрик і аудиту. Використання ATLAS у моделюванні загроз забезпечує сумісну мову зловмисника, усуває прогалини щодо ШІ-векторів і суттєво підвищує валідність тестування захисту ШІ-систем.

CSA MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome). Методологія моделювання загроз для агентних (agentic) ШІ-систем, що заповнює прогалину, яку класичні методики не покривають у частині автономності, емерджентної поведінки та взаємодії багатьох агентів. Представлена у вигляді багаторівневої архітектури з семи рівнів:

1. Foundation Models/Core Services – Базові моделі та основні сервіси;
2. Data Operations – Операції з даними;
3. Agent Frameworks/Application Logic – Фреймворки агентів та прикладна логіка;
4. Deployment & Infrastructure – Розгортання та інфраструктура;
5. Evaluation & Observability – Оцінювання та спостережність;
6. Security & Compliance – Безпека та відповідність вимогам;
7. Agent Ecosystem – Екосистема агентів.

CSA MAESTRO дозволяє декомпозувати ризики та виявляти як внутрішньошарові, так і міжшарові ланцюги атак. Методологія прямо вимагає на етапі опису системи фіксувати цілі агентів, їхні інтерфейси, інструменти, зовнішні API та залежності, а також межі довіри між агентами й підсистемами. На рівні Data Operations MAESTRO очікує на перевірку походження даних, підписів/лінійності, а також контролів проти отруєння, які потім трасуються у сценарії загроз і плани пом'якшення. Для Agent Frameworks підхід підкреслює потребу у тонкому розмежуванні привілеїв агентів (RBAC, токени з обмеженим обсягом, allow-list API), а також протидію ін'єкціям підказок і зловживанню інструментами. Рівень Deployment & Infrastructure зосереджує увагу на CI/CD і контейнерній безпеці, підписуванні образів, захисті від supply-chain-компрометацій і хардингу оркестраторів, що безпосередньо входить до опису ризиків у моделі загроз. Evaluation & Observability вводить вимогу безперервних метрик (дрейф, стабільність, відповідність політикам), що робить моделювання загроз циклічним і доказовим замість одноразового. Шар Security & Compliance примушує пов'язувати результати моделювання загроз з керованими політиками та контролями (на кшталт ISO 27xxx, NIST тощо), забезпечуючи взаємозв'язок з корпоративним ризик-менеджментом. Найвищий шар Agent Ecosystem додає специфічні для багатокомпонентних середовищ ризики (надмірна агентивність, маніпуляція маркетплейсами інструментів тощо), які MAESTRO пропонує моделювати взаємопов'язано з нижчими шарами. CSA прямо порівнює MAESTRO з класичними методами (наприклад STRIDE) і обґрунтовує, чому агентні властивості (недетермінізм, навчання під час виконання, інструментування) потребують розширеної таксономії загроз і шарової декомпозиції. Водночас MAESTRO не замінює ці методи, а надбудовує їх: організація може зберігати прийняту техніку (наприклад, STRIDE) для категоризації загроз, але виконувати аналіз у семи шарах і відслідковувати каскадування впливів. Публікації, що роз'яснюють MAESTRO для індустрій (наприклад, банкінг), додають приклади загроз у кожному шарі й описують контролі (baseline controls), які варто впроваджувати поетапно. У результаті MAESTRO формулює вимоги до моделювання загроз як до багатошарового, профіль-залежного та безперервного процесу з явними артефактами (карта шарів, міжшарові ланцюги, пом'якшення та метрики), що напряму підживлюють дизайн, тестування і моніторинг ШІ-систем.

OWASP GenAI Security Project. Охоплює не лише перелік ризиків, а й повний життєвий цикл безпеки GenAI – від керованості й моделювання загроз до

інцидент-респонсу та інструментів оцінювання. У межах підходу до моделювання загроз проект задає ризик-таксономію (Top-10 2025) як вхідні класи загроз і сценарії зловживань – з акцентом на промпт-ін'єкції (prompt injection), розкриття чутливої інформації, ланцюжків постачання, отруєння даних/моделей, надмірну «агентивність», витік системних промптів і вади RAG/векторних сховищ. Проєкт формує операційні артефакти для практичного застосування при моделюванні загроз: гайд із red team GenAI, що систематизує виявлення вразливостей LLM/агентів у режимі тестування й експлуатації. Для сценаріїв автономних агентів створено Agentic Security Initiative та MAS Threat Modelling Guide, які розширюють класичні методи моделювання загроз під багатошарові/багатоагентні системи і пояснюють, як інтегрувати MAESTRO та відображення на TTPs (ATLAS/ATT&CK). Окремо проект запропонував AI Security Center of Excellence (CoE) Guide, який задає керованість, ролі та процеси довкола моделювання загроз й контролів у великих організаціях. Для підвищення операційної готовності OWASP опублікував GenAI Incident Response Guide 1.0, де класифікацію інцидентів та рівні ШІ-стеку поєднано з фазами IR-процесу, даючи чек-листи й шаблони звітності, які напряму спираються на ризики з Top-10.

Підхід проекту передбачає безперервність: ініціативи та рішення (solution landscape) регулярно поповнюються, зокрема публікуються відгуки про платформи тестування/захисту для «response testing» і «secure code repo», які можна прив’язати до обраної методики моделювання загроз. Вимірюваність і пріоритезацію забезпечує Threat Defense COMPASS 1.0 – легковаговий інструмент/таблиця для оцінки імпакту/ймовірності та вибору пом’якшення, що добре лягає у фазу «Measure/Manage» будь-якої RMF. Загалом проект пропонує методичну зв’язку: таксономія ризиків (Top-10) → практики red-team/response-testing → матриця пріоритезації (COMPASS) → інцидент-респонс (IR Guide), що робить моделювання загроз циклом із доказами ефективності контролів. Для «agentic» сценаріїв керівництва OWASP рекомендують саме шарову декомпозицію (з MAESTRO), бо взаємодія агент-до-агента та емерджентні поведінки породжують крос-шарові ланцюги атак, яких класичні методики не бачать. Публічні матеріали на сайті системно відсилають до міжнародних норм/спільнот (NIST, MITRE) і пропонують канали залучення експертів, завдяки чому артефакти (чек-листи, плейбуки, інструменти) придатні до інтеграції у корпоративні AIMS/RMF. Важливо, що OWASP відкрито розширює обсяг: від розробників/дата-саєнтистів - до CISO/комплаєнс-ролей, щоб моделювання загроз стало корпоративною практикою, а не локальною інженерною активністю. Сукупно OWASP пропонує ризик-інформований і доказовий контур моделювання загроз для GenAI/LLM/агентних систем, який можна «вбудувати» в ISO/IEC-подібні AIMS або NIST RMF-цикли.

Інтегрований підхід моделювання загроз

З огляду на фрагментарність кожної вище описаної методології пропонується інтегрований підхід моделювання загроз у системах ШІ, який поєднує керованість, процесну дисципліну, таксономії атак, архітектурну декомпозицію та операційні артефакти в єдину методику (Рис.1).

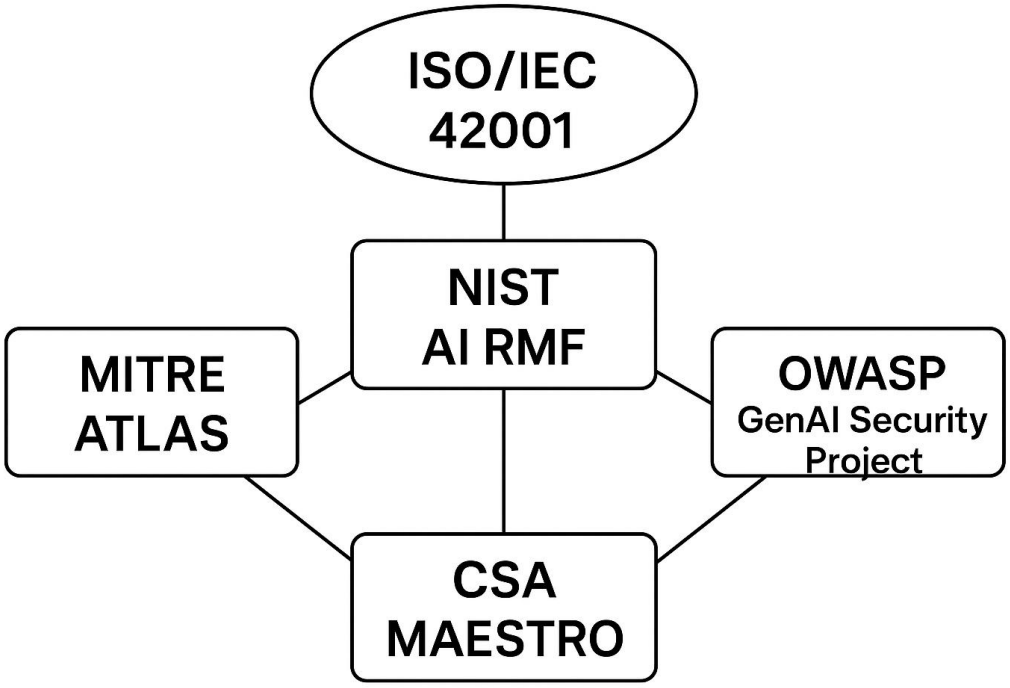

Рис.1. Інтегрований підхід моделювання загроз

ISO/IEC 42001:2023 формує основу керованості через політики, ролі, записи та PDCA-цикл, забезпечуючи доказовість і аудит процесів безпеки. На цьому фундаменті NIST AI RMF 1.0 задає чотири функції – Govern, Map, Measure, Manage – які перетворюють моделювання загроз на безперервний цикл: від визначення контексту й активів до вимірювання ризиків і впровадження контролів. Для наповнення моделей реалістичними сценаріями використовується MITRE ATLAS, що надає тактики, техніки та пом'якшення (TTPs) на основі реальних атак, а також успішні приклади для перевірки гіпотез. Архітектурну площину описується CSA MAESTRO, на основі семи рівневої структури (Foundation Models, Data Ops, Agent Frameworks, Deployment, Observability, Security & Compliance, Ecosystem) для аналізу міжрівневих залежностей і каскадних загроз. Операційний рівень забезпечує OWASP GenAI Security Project, що надає таксономію ризиків (Top-10), гайд із red-teaming, інструмент пріоритезації COMPASS та плейбуки інцидент-респонсу.

Інтеграція цих компонентів (Таблиця 2) дозволяє почати з відображення системи за MAESTRO, підсилити сценарії загроз техніками ATLAS, застосувати чек-листи OWASP для перевірки контролів і завершити цикл керованістю ISO та RMF. Такий підхід робить моделювання загроз не разовою вправою, а керованим процесом із доказами ефективності, метриками та аудитом. У цьому підході враховуються як технічні, так і соціотехнічні ризики, включаючи вплив на користувачів і суспільство. Важливо, що інтегрований підхід підтримує профільну адаптацію: для LLM, агентних систем або RAG-ланцюгів можна формувати окремі карти загроз і контролів. Поєднання ISO та RMF гарантує узгодженість із міжнародними нормами, а ATLAS і OWASP забезпечують актуальність сценаріїв і практичність перевірок. MAESTRO додає прозорість у складних архітектурах, де загрози виникають на стиках шарів.

Таблиця 2

Роль методів у інтегрованому підході моделювання загроз

Стандарт / Методологія	Роль	Місце у моделюванні загроз
ISO/IEC 42001:2023	Керованість і аудит	Формалізує політики, ролі, PDCA-цикл; забезпечує доказовість, трасовність і сертифікаційну перевірку
NIST AI RMF 1.0	Методологія процесу	Задає цикл Govern–Map–Measure–Manage; інтегрує моделювання загроз у життєвий цикл ШІ.
MITRE ATLAS	Таксономія атак	Надає TTPs (тактики, техніки, процедури) та кейси для побудови реалістичних сценаріїв загроз
CSA MAESTRO	Архітектурна декомпозиція	Пропонує семишарову модель для аналізу міжрівневих залежностей і каскадних атак
OWASP GenAI Security Project	Операційні артефакти та практичні інструменти	Надає Top-10 ризиків, гайд з red-teaming, COMPASS для пріоритезації та плейбуки інцидент-респонсу

ВИСНОВКИ НА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Сучасні методи моделювання загроз для систем штучного інтелекту залишаються фрагментарними та орієнтованими на окремі аспекти безпеки. Жоден із розглянутих підходів не забезпечує повного покриття життєвого циклу ШІ, що створює потребу в інтегрованій методиці. Стандарти ISO/IEC 42001 та NIST AI RMF формують основу керованості та процесної дисципліни, але не деталізують технічні сценарії атак. MITRE ATLAS надає таксономію реальних TTPs, проте потребує поєднання з управлінськими вимогами для забезпечення доказовості. CSA MAESTRO пропонує багатошарову декомпозицію ризиків, що є критично важливою для агентних систем, але вимагає інтеграції з профілями ризику. OWASP GenAI Security Project доповнює підхід операційними артефактами, чек-листами та інструментами пріоритезації, проте не охоплює стратегічний рівень керованості. З огляду на це доцільним є формування єдиного підходу, який поєднує керованість, процесний цикл, таксономії атак та практичні інструменти.

Інтегрований підхід передбачає поєднання цих компонентів у єдину методику моделювання загроз. У результаті організація отримує методику, яка охоплює весь життєвий цикл ШІ, від дизайну до експлуатації та виведення з роботи. Такий підхід підвищує стійкість систем, скорочує час реагування на інциденти та знижує ризики, пов’язані з емерджентною поведінкою моделей. Він також створює основу для автоматизації: карти шарів, матриці технік і чек-листи, які можна інтегрувати у CI/CD-процеси. Інтегрований підхід забезпечує не лише технічну безпеку, а й довіру до ШІ, що є ключовим фактором для його масштабного впровадження.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Neretin, O., & Kharchenko, V. (2022). Ensuring cybersecurity of artificial intelligence systems: Analysis of vulnerabilities, attacks, and countermeasures. *Bulletin of the National University "Lviv Polytechnic". Information Systems and Networks, (12)*, 7–22. http://nbuv.gov.ua/UJRN/VNULPICM_2022_12_4
2. Dudykovych, V. B., Mykytyn, H. V., & Kret, T. B. (2015). Multilevel intelligent control systems: Guarantee capability and object security. *Information Processing Systems, (4)*, 92–95. http://nbuv.gov.ua/j-pdf/soi_2015_4_21.pdf
3. Martseniuk, Ye. V., Partyka, A. I., & Kret, T. B. (2025). Study of artificial intelligence vulnerabilities and development of a comprehensive organizational security model. *Modern Information Protection, 1*(61), 206–218. https://doi.org/10.31673/2409-7292.2025.018929
4. National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1
5. Cloud Security Alliance. (2025, February 6). *Agentic AI Threat Modeling Framework: MAESTRO.* https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro
6. MITRE Corporation. (n.d.). *ATLAS Matrix.* https://atlas.mitre.org/matrices/ATLAS
7. OWASP Foundation. (n.d.). *Gen AI Security Project: Introduction and background.* https://genai.owasp.org/introduction-genai-security-project
8. Amazon Web Services (AWS). (n.d.). *How to approach threat modeling.* https://aws.amazon.com/blogs/security/how-to-approach-threat-modeling
9. Amazon Web Services (AWS). (n.d.). *Threat modeling your generative AI workload to evaluate security risk.* https://aws.amazon.com/blogs/security/threat-modeling-your-generative-ai-workload-to-evaluate-security-risk
10. Yevseiev, S. P., Shmatko, O. V., Akhiezer, O. B., Sokol, V. Ye., & Chernova, N. L. (2025). *Attacks on artificial intelligence systems: Educational and practical manual* (S. P. Yevseiev, Ed.). Kharkiv: NTU "KhPI"; Lviv: Novyi Svit-2000.
11. Straiker AI. (2025). *Comparing AI security frameworks: OWASP, CSA, NIST, and MITRE.* https://www.straiker.ai/blog/comparing-ai-security-frameworks-owasp-csa-nist-and-mitre
12. International Organization for Standardization (ISO). (2023). *ISO/IEC 42001:2023 — Information technology – Artificial intelligence – Management system.* https://www.iso.org/standard/42001
13. Microsoft. (n.d.). *The STRIDE Threat Model.* https://learn.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20)
14. Mauri, L., & Damiani, E. (2022). Modeling threats to AI-ML systems using STRIDE. *Sensors, 22*(17), 6662. https://doi.org/10.3390/s22176662
15. Khan, R., Sarkar, S., Mahata, S. K., & Jose, E. (2024). Security threats in agentic AI systems. *arXiv preprint arXiv:2410.14728.* https://doi.org/10.48550/arXiv.2410.14728

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (30), 2025
Спеціальний випуск ISSN 2663 – 4023
Special issue

Kret Taras
Senior Lecturer, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0002-6333-3190
taras.b.kret@lpnu.ua

Martseniuk Yevhenii
Assistant, Department of Information Security
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0009-2289-0968
yevhenii.v.martseniuk@lpnu.ua

INTEGRATED APPROACH TO THREAT MODELING IN ARTIFICIAL INTELLIGENCE SYSTEMS

Abstract. This paper substantiates the relevance of threat modeling for artificial intelligence (AI) systems in the context of increasing model autonomy and the emergence of new attack vectors. It demonstrates that traditional methods fail to account for the specific nature of AI, creating the need for a comprehensive approach capable of covering the entire system lifecycle. The methodological foundation of the integrated approach combines international standards and industry best practices: ISO/IEC 42001:2023 ensures governance and auditing, NIST AI RMF 1.0 defines the process cycle Govern–Map–Measure–Manage, MITRE ATLAS enriches models with realistic attack scenarios, CSA MAESTRO introduces multi-layer architectural decomposition, and OWASP GenAI Security Project provides operational artifacts and prioritization tools. This synthesis enables the integration of strategic policies, technical taxonomies, and practical playbooks into a single managed process. The proposed approach makes threat modeling continuous and evidence-based, ensuring traceability from threat identification to control implementation and performance metrics. It addresses both technical and socio-technical risks, including impacts on users and society, and supports profile-specific adaptation for various system types—from LLMs to agent-based platforms. Integration with CI/CD pipelines and automation of security checks improves response speed and reduces security costs. The scientific novelty lies in forming a holistic vision that combines governance, process discipline, architectural analysis, and operational instruments. The practical significance is in the ability to apply this approach to develop comprehensive protection strategies aligned with international standards and suitable for certification audits. The integrated approach establishes a foundation for large-scale AI deployment with proven security and trust during threat modeling. It not only enhances system resilience but also creates a standardized risk management framework that meets modern cybersecurity challenges.

Keywords: artificial intelligence; threat modeling; ISO/IEC 42001; NIST AI RMF; MITRE ATLAS; CSA MAESTRO; OWASP GenAI; integrated approach; risk management; AI security.

REFERENCES

1. Neretin, O., & Kharchenko, V. (2022). Ensuring cybersecurity of artificial intelligence systems: Analysis of vulnerabilities, attacks, and countermeasures. *Bulletin of the National University "Lviv Polytechnic". Information Systems and Networks, (12)*, 7–22. http://nbuv.gov.ua/UJRN/VNULPICM_2022_12_4
2. Dudykovych, V. B., Mykytyn, H. V., & Kret, T. B. (2015). Multilevel intelligent control systems: Guarantee capability and object security. *Information Processing Systems, (4)*, 92–95. http://nbuv.gov.ua/j-pdf/soi_2015_4_21.pdf
3. Martseniuk, Ye. V., Partyka, A. I., & Kret, T. B. (2025). Study of artificial intelligence vulnerabilities and development of a comprehensive organizational security model. *Modern Information Protection, 1*(61), 206–218. https://doi.org/10.31673/2409-7292.2025.018929
4. National Institute of Standards and Technology (NIST). (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

5. Cloud Security Alliance. (2025, February 6). *Agentic AI Threat Modeling Framework: MAESTRO.* https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro
6. MITRE Corporation. (n.d.). *ATLAS Matrix.* https://atlas.mitre.org/matrices/ATLAS
7. OWASP Foundation. (n.d.). *Gen AI Security Project: Introduction and background.* https://genai.owasp.org/introduction-genai-security-project
8. Amazon Web Services (AWS). (n.d.). *How to approach threat modeling.* https://aws.amazon.com/blogs/security/how-to-approach-threat-modeling
9. Amazon Web Services (AWS). (n.d.). *Threat modeling your generative AI workload to evaluate security risk.* https://aws.amazon.com/blogs/security/threat-modeling-your-generative-ai-workload-to-evaluate-security-risk
10. Yevseiev, S. P., Shmatko, O. V., Akhiezer, O. B., Sokol, V. Ye., & Chernova, N. L. (2025). *Attacks on artificial intelligence systems: Educational and practical manual* (S. P. Yevseiev, Ed.). Kharkiv: NTU "KhPI"; Lviv: Novyi Svit-2000.
11. Straiker AI. (2025). *Comparing AI security frameworks: OWASP, CSA, NIST, and MITRE.* https://www.straiker.ai/blog/comparing-ai-security-frameworks-owasp-csa-nist-and-mitre
12. International Organization for Standardization (ISO). (2023). *ISO/IEC 42001:2023 — Information technology – Artificial intelligence – Management system.* https://www.iso.org/standard/42001
13. Microsoft. (n.d.). *The STRIDE Threat Model.* https://learn.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20)
14. Mauri, L., & Damiani, E. (2022). Modeling threats to AI-ML systems using STRIDE. *Sensors, 22*(17), 6662. https://doi.org/10.3390/s22176662
15. Khan, R., Sarkar, S., Mahata, S. K., & Jose, E. (2024). Security threats in agentic AI systems. *arXiv preprint arXiv:2410.14728.* https://doi.org/10.48550/arXiv.2410.14728