# AI Security — новий безпековий тренд чи необхідність?

**Костянтин ОНІЩЕНКО, Product Manager компанії Oberig IT, розповідає про загрози, пов'язані з використанням штучного інтелекту, і про те, як від них захиститися.**

У 2026 році штучний інтелект (AI) став невід'ємною частиною бізнесу й повсякденного життя. AI використовують для пошуку, аналізу, генерації даних та автоматизації процесів, і цей тренд лише зростає. Для IT-команд це створює нові виклики: масштабування інфраструктури, вибір між власним ЦОД і хмарою, контроль даних у публічних AI-сервісах та управління власними LLM/ML-моделями. У результаті IT-периметр стає значно складнішим і охоплює різні сервіси, платформи та середовища розміщення.

## AI та сучасні виклики безпеки

Команди IT і кібербезпеки розуміють, що впровадження AI-інструментів і розширення інфраструктури потребують оцінювання ризиків, контролю безпеки та механізмів реагування. Ми проаналізували AI-загрози за 2025 рік на основі звітів некомерційних організацій OWASP, NIST та International AI Safety Report, а також досліджень компаній у сфері моделювання атак, таких як Adversa AI, Cymulate і DeepStrike.

Дослідження показують, що близько 70% кіберінцидентів пов'язані з Generative AI. Зі свого боку, Agentic AI створює ще більш критичні ризики через автономні дії, які впливають на фінанси та бізнес-процеси. При цьому 35% інцидентів спричинені простими атаками типу prompt injection, і загалом більшість атак не потребують malware — достатньо взаємодії через текстові запити (prompts). Основні типи ризиків, що формують приблизно 70% випадків, — це prompt injection, unsafe outputs і data leakage, що вказує на системну проблему маніпуляції вхідними даними, неконтрольованих відповідей і витоків даних.

**Prompt injection** дозволяє приховано змінювати поведінку LLM через prompts або RAG-контент, змушуючи систему виконувати неавторизовані дії, тоді як **data leakage** призводить до розкриття чутливої інформації з документів або історії запитів, а **unsafe outputs** — до генерації неправдивих або небезпечних рішень і дій AI-агентів. Усі ці ризики потребують поєднання runtime-контролю, політик безпеки, ізоляції даних і перевірки критичних дій людиною перед їх виконанням.

## Захист AI — один із фокусних напрямків безпеки навіть серед найдосвідченіших гравців ринку

**2016:** AI-чатбот Microsoft Tay був скомпрометований через 16 годин після запуску через онлайн-маніпуляції.
**2023:** Samsung обмежила використання ChatGPT після витоку внутрішнього вихідного коду співробітниками.
**2024:** AI-бібліотеку Ultralytics скомпрометували для розповсюдження криптомайнера через PyPI.
**2025:** Amazon Q під час supply chain-інциденту наражає близько 950 тис. розробників на шкідливі команди.

## Сучасні підходи до контролю ризиків, пов'язаних з AI

Світові організації у сфері аналізу і регламентування кібербезпеки вже активно формують і оновлюють стандарти та рекомендації для захисту AI-систем.

NIST розробила AI Risk Management Framework (AI RMF) для безпечного проєктування та експлуатації AI. OWASP досліджує ризики та вразливості AI у проєктах OWASP Top 10 for LLMs і OWASP AI Exchange. Європейський Союз ухвалив EU AI Act — перший комплексний закон для регулювання AI, який визначає вимоги до безпеки, прозорості та контролю ризиків. Також завершується затвердження стандарту ISO/IEC FDIS 27090, який стане частиною екосистеми ISO 27001 для безпеки AI.

> *ШІ більше не просто трендовий інструмент, а частина інфраструктури, яка потребує детальної уваги до проєктування, розроблення та захисту від загроз.*

## Чому поширені підходи до безпеки ШІ призводять до перебоїв у продуктивному середовищі

Більшість AI Security-рішень повторюють помилки хмарної безпеки: фокусуються на окремих інструментах замість єдиної операційної моделі. У результаті виникають фрагментована відповідальність, перевантажений SOC і відсутність цілісної картини ризиків.

Проблеми зазвичай проявляються у трьох підходах. AppSec охоплює код і CI/CD, але не контролює поведінку моделей, prompts і tool-виклики,

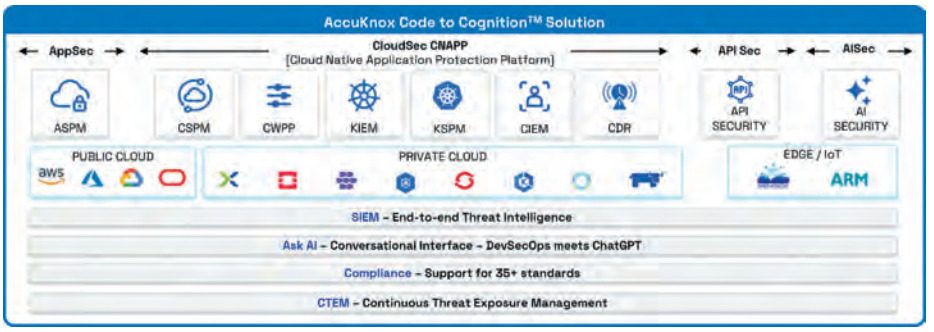
через що runtime-атаки залишаються непоміченими. CloudSec виявляє конфігураційні помилки, але не бачить AI-специфічні атаки, витоки даних і зловживання агентами, що створює надлишок сповіщень без ефективного реагування. GRC (Governance, Risk, and Compliance) забезпечує політики та відповідність, але без runtime-контролю аудит стає неповним, а відповідальність — розмитою.

## Якого підходу вимагає сучасне AI Security рішення?

Ефективна AI-безпека базується не на боротьбі з кожною новою AI-атакою, а на застосуванні перевірених принципів кібербезпеки до AI-систем: мінімальні привілеї, контроль змін і постійний runtime-контроль. Ключовими елементами такого підходу є інвентаризація LLM/ML-моделей, агентів і AI-сервісів, контроль даних і доступу, захист prompt/response, моніторинг поведінки AI у runtime, постійна перевірка безпеки через red teaming та regression testing, а також контроль відповідності й аудит відповідно до вимог NIST AI RMF та EU AI Act.

## Наш підхід до вибору рішення

Наприкінці 2025 року команда Oberig IT проаналізувала сучасні AI-загрози та підходи до безпечної розробки й експлуатації AI-систем і розпочала пошук AI Security-рішення для потреб українських компаній. Ключовими вимогами були підтримка сучасних підходів AI Security: інтеграції з CI/CD, MLOps і класичними security-інструментами, контролю AI на всіх етапах lifecycle, а також можливості роботи як у власному ЦОД, так і в публічних чи приватних хмарах. Важливими були досвід виробника у великих AI та безпекових проєктах, а також гнучкість у адаптації рішень для клієнтів. У результаті наприкінці 2025 року ми обрали виробника ***AccuKnox*** як стратегічного партнера для захисту хмарних, гібридних та AI-середовищ.




## Чому AccuKnox?

Рішення AccuKnox розробляються у активній співпраці з інститутом SRI International, відомим участю у створенні Siri та Nuance. Рішення поєднує підходи AI Security з технологіями open source, зокрема KubeArmor, що дозволяє не лише виявляти, а й запобігати загрозам у runtime на різних рівнях інфраструктури.

Підхід AccuKnox — це не окремий інструмент, а комплексна модель захисту AI та хмарної інфраструктури, що відображається в архітектурі платформи та гнучкому ліцензуванні, яке дозволяє впроваджувати лише потрібні функції залежно від задач і зрілості AI-проєктів.

Можливі сценарії використання:
Якщо організація зосереджена на розробленні коду, AccuKnox забезпечує **захист CI/CD-процесів** і дозволяє інтегрувати безпеку безпосередньо в операційну модель та інструменти розроблення.

**Для експлуатації** AccuKnox пропонує захист Kubernetes та інфраструктури з безперервним моніторингом конфігурацій і стану безпеки робочого середовища.

**Для публічних і приватних LLM** AccuKnox надає інструменти аналізу моделей і контролю запитів/відповідей AI, щоб запобігати витокам, компрометації даних і несанкціонованому доступу до AI.

## Підсумки

AI-безпека вже є питанням рівня топменеджменту багатьох компаній, оскільки AI інтегрований у критичні бізнес-процеси й може виконувати дії в реальному середовищі. Надійна і ефективна модель безпеки — це не набір окремих рішень, а поєднання моделі Zero Trust, безперервної перевірки, runtime-захисту та комплаєнс-контролів, що працюють як єдина система протягом усього життєвого циклу AI.