Zastosowanie AI w Zadaniach Cyberbezpieczeństwa

Wykład 1

Ramy cyberbezpieczeństwa · NIST AI RMF · Bezpieczeństwo modeli AI

Plan wykładu

4 główne bloki tematyczne

01 **Budowanie ram cyberbezpieczeństwa**

NIST CSF 2.0, Triada CIA

02 **NIST AI Risk Management Framework**

Zasady, odbiorcy, AI RMF Core

03 **Jak zabezpieczyć modele AI?**

Dane, modele, użytkowanie

04 **Zagrożenia i środki zaradcze**

Poisoning, injection, guardrails

01

NIST Cybersecurity Framework

Budowanie ram bezpieczeństwa cybernetycznego

NIST Cybersecurity Framework 2.0

Sześć kluczowych funkcji

GOVERN

Kontekst organizacji, tolerancja ryzyka, misja i cele strategiczne

IDENTIFY

Inwentaryzacja zasobów, ocena ryzyka, luki bezpieczeństwa

PROTECT

Kontrola dostępu, szkolenia, zabezpieczenia techniczne

DETECT

Monitoring, wykrywanie anomalii, alerty i korelacja zdarzeń

RESPOND

Reagowanie na incydenty, komunikacja, izolacja zagrożeń

RECOVER

Przywracanie działania, wnioski, doskonalenie procesów

Triada CIA

Fundamenty bezpieczeństwa informacji

Confidentiality

Poufność

Dane dostępne tylko dla uprawnionych. Mechanizmy: szyfrowanie, kontrola dostępu, klasyfikacja danych.

Integrity

Integralność

Dane są dokładne i niezmienione bez autoryzacji. Mechanizmy: sumy kontrolne, podpisy cyfrowe, logi audytowe.

Availability

Dostępność

Systemy działają, kiedy są potrzebne. Mechanizmy: redundancja, DDoS protection, backup & recovery.

02

NIST AI Risk Management Framework

Zarządzanie ryzykiem sztucznej inteligencji

NIST AI RMF – czym jest?

Dobrowolne narzędzie zarządzania ryzykiem AI

Definicja

Dobrowolne narzędzie wspierające organizacje w projektowaniu, tworzeniu, wdrażaniu oraz wykorzystywaniu systemów AI — w celu zarządzania ryzykiem i promowania godnej zaufania, odpowiedzialnej sztucznej inteligencji.

Kto jest odbiorcą?

Firmy technologiczne

Projektujące i rozwijające systemy AI

Sektor publiczny & biznes

Instytucje wdrażające AI: finanse, zdrowie, administracja

Użytkownicy końcowi

Korzystający z rozwiązań opartych na AI

Audytorzy & compliance

Zespoły odpowiedzialne za zgodność z przepisami

Nauka & edukacja

Środowisko badawcze i akademickie

6 zasad zaufanej i odpowiedzialnej AI

NIST AI RMF Core Principles

Praworządność

Systemy AI zgodne z prawem i regulacjami prawnymi

Uczciwość

Eliminowanie uprzedzeń, sprawiedliwe traktowanie wszystkich grup

Przejrzystość

Otwartość co do działania systemów AI i podejmowanych decyzji

Bezpieczeństwo

Odporność na ataki, niezawodność i cyberbezpieczeństwo AI

Prywatność

Ochrona danych osobowych w całym cyklu życia systemu AI

Odpowiedzialność

Jasne mechanizmy rozliczalności i nadzoru nad AI

AI RMF Core – 4 kluczowe funkcje

Kultura zarządzania ryzykiem AI

GOVERN

Ustalenie kultury zarządzania ryzykiem AI – polityki, role, odpowiedzialność i procesy organizacyjne.

MAP

Identyfikacja i klasyfikacja ryzyk AI – kontekst, interesariusze, potencjalne skutki.

MEASURE

Analiza i ocena ryzyk – metryki, ewaluacja, testy bezpieczeństwa modeli.

MANAGE

Priorytetyzacja i reagowanie na ryzyki – plany mitygacji, monitoring, doskonalenie.

Źródło: https://www.nist.gov/itl/ai-risk-management-framework

03

Jak zabezpieczyć modele AI?

Dane · Model · Użytkowanie

Jak działa generatywna AI – punkty krytyczne

Trzy warstwy wymagające zabezpieczenia

01

Dane

Dane treningowe, zestawy danych, dane wejściowe użytkowników.
Zagrożenia: poisoning, exfiltration, leakage.

02

Model

Architektura, wagi, interfejsy API.
Zagrożenia: tylne drzwi, naruszenia IP, niepewne źródła.

03

Użytkowanie

Interakcja użytkowników, endpointy, wyniki modelu.
Zagrożenia: prompt injection, DoS, kradzież modelu.

Zabezpieczenia danych

Ochrona na poziomie danych treningowych i wejściowych

Najpoważniejsze zagrożenia

- Zatrucie danych (data poisoning)
- Kradzież danych (exfiltration)
- Wyciek danych (leakage)

Jak się bronić?

- Identyfikacja i klasyfikacja danych
- Szyfrowanie (at rest & in transit)
- Kontrola dostępu i monitorowanie

Bezpieczeństwo modeli AI

Zagrożenia i środki zaradcze na poziomie modelu

Najpoważniejsze zagrożenia

- Niepewne źródło modelu
- Model jako "tylne drzwi" (backdoor)
- Wtyczki i API bez autoryzacji
- Naruszenia praw autorskich (IP)

Jak się bronić?

- Weryfikacja źródeł modeli
- Skany antywirusowe i checksumy
- Wzmacnianie systemów (hardening)
- Kontrola uprawnień i zgodność z prawem

04

Jak zabezpieczyć modele AI?

Dane · Model · Użytkowanie

Zabezpieczenia użytkownika

Zagrożenia i ochrona na poziomie interakcji z modelem

Najpoważniejsze zagrożenia

- Prompt injection
- Atak DoS (Denial of Service)
- Kradzież modelu (model extraction)

Jak się bronić?

- Monitorowanie i analiza zapytań
- Guardrails – ograniczenia wejść/wyjść
- Systemy SIEM i SOAR dla AI
- Narzędzia AI security (np. LLM Guard)

Dziękuję za uwagę

AI zmienia krajobraz zagrożeń

Nowe wektory ataku wymagają nowych strategii ochrony

NIST AI RMF to fundament

Dobrowolne, ale niezbędne ramy zarządzania ryzykiem AI

Bezpieczeństwo = Dane + Model + Użytkowanie

Każda warstwa wymaga osobnego podejścia i narzędzi