SYLABUS

DOTYCZY CYKLU KSZTAŁCENIA 2026/2027 - 2029/2030

Rok akademicki 2028/2029

1. PODSTAWOWE INFORMACJE O PRZEDMIOCIE

Nazwa przedmiotu	*bezpieczeństwo AI*
Kod przedmiotu*	
Nazwa jednostki prowadzącej kierunek	*Instytut Informatyki, Wydział Nauk Ścisłych i Technicznych*
Nazwa jednostki realizującej przedmiot	*Instytut Informatyki, Wydział Nauk Ścisłych i Technicznych*
Kierunek studiów	*sztuczna inteligencja*
Poziom studiów	*studia I stopnia*
Profil	*ogólnoakademicki*
Forma studiów	*stacjonarne*
Rok i semestr/y studiów	*rok III, semestr 5*
Rodzaj przedmiotu	*przedmiot kierunkowy*
Język wykładowy	*polski*
Koordynator	*dr inż. Marcin Ochab*
Imię i nazwisko osoby prowadzącej / osób prowadzących	*dr inż. Marcin Ochab*

***** - *opcjonalnie, zgodnie z ustaleniami w Jednostce*

1.1.Formy zajęć dydaktycznych, wymiar godzin i punktów ECTS

Semestr (nr)	Wykł.	Ćw.	Konw.	Lab.	Sem.	ZP	Prakt.	Inne (jakie?)	**Liczba pkt. ECTS**
5	15			15					2

1.2. Sposób realizacji zajęć

zajęcia w formie tradycyjnej

1.3 Forma zaliczenia przedmiotu (z toku)

wykład – zaliczenie bez oceny, laboratoria – zaliczenie z oceną

2.WYMAGANIA WSTĘPNE

podstawy uczenia maszynowego (regresja, klasyfikacja, overfitting), podstawy bezpieczeństwa informacji (podstawowe ataki sieciowe, OWASP Top 10), podstawy programowania

3. CELE, EFEKTY UCZENIA SIĘ, TREŚCI PROGRAMOWE I STOSOWANE METODY DYDAKTYCZNE

3.1 Cele przedmiotu

C1	Rozumieć modele zagrożeń dla systemów AI (klasyczne ML, deep learning, LLM).
C2	Umieć klasyfikować i analizować ataki na systemy uczenia maszynowego (poisoning, evasion, inference).
C3	Znać metody obrony (robust training, monitoring, izolacja danych, hardening pipeline'u ML).
C4	Rozumieć aspekty prawne i etyczne, w tym odpowiedzialność za szkody wywołane działaniem AI.
C5	Potrafić zaprojektować prosty, odporny na ataki pipeline ML/AI i udokumentować przyjęty model zagrożeń.

3.2 Efekty uczenia się dla przedmiotu

EK (efekt uczenia się)	Treść efektu uczenia się zdefiniowanego dla przedmiotu	Odniesienie do efektów kierunkowych [1]
EK_01	Zna typowe wektory ataku na modele ML i systemy oparte o AI. Rozumie zależności między jakością danych, prywatnością, a bezpieczeństwem modeli.	K_W09
EK_02	Zna aktualne ramy regulacyjne (AI Act UE, RODO w kontekście danych treningowych, wytyczne dot. systemów wysokiego ryzyka).	K_W12
EK_03	Potrafi przeprowadzić podstawową analizę bezpieczeństwa systemu AI (model zagrożeń, diagramy przepływu danych). Umie zademonstrować wybrane ataki (np. adversarial examples, membership inference w warunkach lab). Potrafi zaprojektować i wdrożyć podstawowe mechanizmy obrony oraz monitorowania nadużyć.	K_U14
EK_04	Ma świadomość odpowiedzialności etycznej przy projektowaniu i atakowaniu systemów AI.	K_K01

3.3 Treści programowe

A. Problematyka wykładu

Wprowadzenie do bezpieczeństwa AI
Modele zagrożeń dla systemów AI
Bezpieczny cykl życia systemu AI
Ataki na dane treningowe – data poisoning, Obrona przed poisoningiem
Ataki evasion (adversarial examples) i obrona przed nimi
Prywatność a AI, techniki ochrony prywatności
Bezpieczeństwo dużych modeli językowych

[1] W przypadku ścieżki kształcenia prowadzącej do uzyskania kwalifikacji nauczycielskich uwzględnić również efekty uczenia się ze standardów kształcenia przygotowującego do wykonywania zawodu nauczyciela.

Etyka, odpowiedzialność i regulacje

B. Problematyka laboratoriów

Analiza architektury prostego systemu ML, identyfikacja zasobów i wektorów ataku.
Wstrzyknięcie prostego poisoning w klasyfikatorze
Generowanie adversarial examples na pretrenowanym modelu i implementacja obrony przed nimi.
Scenariusze ataków na LLM: prompt injection, data exfiltration w środowisku odizolowanym od danych realnych.
Praca projektowa w zespołach – zaprojektowanie i częściowa implementacja pipeline'u AI z elementami zabezpieczenia i monitoringiem nadużyć.

3.4 Metody dydaktyczne

Wykład: wykład z prezentacją multimedialną.
Laboratorium: wykonywanie zadań praktycznych przy komputerze, projekt.

4. METODY I KRYTERIA OCENY

4.1 Sposoby weryfikacji efektów uczenia się

Symbol efektu	Metody oceny efektów uczenia się (np.: kolokwium, egzamin ustny, egzamin pisemny, projekt, sprawozdanie, obserwacja w trakcie zajęć)	Forma zajęć dydaktycznych (w, ćw, ...)
EK_01, EK_02	kolokwium	wykład
EK_03	projekt	laboratorium
EK_04	obserwacja w trakcie zajęć	laboratorium

4.2 Warunki zaliczenia przedmiotu (kryteria oceniania)

Zaliczenie wykładu uzyskiwane jest na podstawie kolokwium, weryfikującego znajomość zagadnień przedstawionych w ramach wykładu.
Efekty EK_01 i EK_02 uznaje się za zaliczone, gdy udzielone zostanie co najmniej 50% prawidłowych odpowiedzi na pytania przyporządkowane do każdego z nich.
Zaliczenie wykładu następuje, gdy EK_01 i EK_02 są zaliczone na „zal".

Zaliczenie laboratorium następuje na podstawie wykonanego projektu (w ramach zespołów), dotyczącego wybranego tematu z zagadnień omawianych na laboratorium. Za projekt wystawiana jest ocena w skali 2.0 - 5.0 proporcjonalnie do uzyskanej liczby punktów, możliwych do zdobycia za wykonanie praktycznej części projektu oraz za dokumentację (opisującą zrealizowany projekt).

Efekt EK_03 jest uznany za zaliczony, gdy projekt wykonany przez zespół uzyska przynajmniej 50% możliwych punktów.

Aktywność studenta podczas zajęć jest oceniania w ramach efektu EK_04.

Ocena końcowa z laboratorium jest wystawiana na postawie oceny za efekt EK_03, pod warunkiem, że efekt EK_04 został zaliczony na 'zal'.

5. CAŁKOWITY NAKŁAD PRACY STUDENTA POTRZEBNY DO OSIĄGNIĘCIA ZAŁOŻONYCH EFEKTÓW W GODZINACH ORAZ PUNKTACH ECTS

Forma aktywności	Średnia liczba godzin na zrealizowanie aktywności
Godziny z harmonogramu studiów	30
Inne z udziałem nauczyciela akademickiego (udział w konsultacjach, egzaminie)	-
Godziny niekontaktowe – praca własna studenta (przygotowanie do zajęć, egzaminu, napisanie referatu itp.)	20
SUMA GODZIN	50
SUMARYCZNA LICZBA PUNKTÓW ECTS	**2**

** Należy uwzględnić, że 1 pkt ECTS odpowiada 25-30 godzin całkowitego nakładu pracy studenta.*

6. PRAKTYKI ZAWODOWE W RAMACH PRZEDMIOTU

wymiar godzinowy	–
zasady i formy odbywania praktyk	–

7. LITERATURA

Literatura podstawowa:

1. Machine learning, Python i data science: wprowadzenie / Andreas C. Müller, Sarah Guido; przekład: Michał Sternik. - Gliwice: Helion, © 2021.
2. Jak projektować systemy uczenia maszynowego: iteracyjne tworzenie aplikacji gotowych do pracy / Chip Huyen ; przekład: Jacek Janusz. - Gliwice: Helion, © 2023.
3. Sztuczna inteligencja i uczenie maszynowe dla programistów: praktyczny przewodnik po sztucznej inteligencji / Laurence Moroney ; przekład: Jacek Janusz. - Gliwice: Helion, © 2021.

Literatura uzupełniająca:

1. Adversarial Machine Learning (Yevgeniy Vorobeychik, Murat Kantarcioglu), Morgan & Claypool Publishers, 2018
2. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (NIST.AI.100-2e2025, Apostol Vassilev et al.), NIST, 2025.