

To wciąż tylko maszyna

Kwestie bezpieczeństwa w korzystaniu ze sztucznej inteligencji

Warsztat #7



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Agenda

1. Powitanie i wprowadzenie
2. Wprowadzenie do sztucznej inteligencji i kwestii bezpieczeństwa
3. Dlaczego sztucznej inteligencji nie należy ufać bezwarunkowo?
4. Typowe zagrożenia bezpieczeństwa w sztucznej inteligencji
5. Narzędzia online do zabezpieczania sztucznej inteligencji
6. Sesja praktyczna: Granie w gry typu Prompt Injection
7. Sesja praktyczna: Odkrywanie bezpiecznych środowisk piaskownicy AI
8. Aplikacje w świecie rzeczywistym
9. Pytania i odpowiedzi oraz dyskusja



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Wprowadzenie do sztucznej inteligencji i kwestii bezpieczeństwa



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Zrozumienie i ograniczanie ryzyka

Sztuczna inteligencja jest potężna, ale może być ryzykowna, jeśli nie jest odpowiednio zabezpieczona

- Systemy sztucznej inteligencji popełniające błędy z powodu sfałszowanych danych
- Naruszenia prywatnych informacji
- Przedstawianie błędnych odpowiedzi w przekonujący sposób



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Dlaczego nie należy bezwarunkowo ufać sztucznej inteligencji?

- **Odcięcie wiedzy:** Sztuczna inteligencja może nie mieć najnowszych informacji
- **Źli aktorzy:** Złośliwa manipulacja wynikami sztucznej inteligencji
- **Uprzedzenia:** Sztuczna inteligencja może dziedziczyć i wzmacniać uprzedzenia z danych szkoleniowych.
- **Zatruwanie danych treningowych:** Korupcja danych wykorzystywanych do uczenia AI
- **Modele open source vs. modele własnościowe:**
 - **Open Source:** Większa przejrzystość, ale potencjalnie więcej luk w zabezpieczeniach
 - **Modele własnościowe:** Mniejsza przejrzystość, ale potencjalnie lepsza kontrola bezpieczeństwa



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Typowe zagrożenia bezpieczeństwa w sztucznej inteligencji



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Typowe zagrożenia bezpieczeństwa w sztucznej inteligencji

- **Ataki przeciwników:** Zmusza sztuczną inteligencję do popełniania błędów
- **Zatruwanie danych:** Przekazywanie sztucznej inteligencji złych informacji
- **Inwersja modelu:** Wyciąganie wrażliwych danych
- **Naruszenia prywatności:** Nieautoryzowany dostęp do danych



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Ćwiczenie praktyczne: Granie w gry typu Prompt Injection



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Granie w gry typu Prompt Injection

Gandalf's Spellbook

- Odwiedź stronę z grą typu prompt injection: Gandalf's Spellbook
- Postępuj zgodnie z instrukcjami i celowo wchodź w interakcję ze sztuczną inteligencją, aby odkryć luki w zabezpieczeniach.
- Porównaj wyniki z innymi graczami i omów wyniki



<https://gandalf.lakera.ai>



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Granie w gry typu Prompt Injection

TensorTrust AI

- Odwiedź stronę z grą typu prompt injection: TensorTrust AI
- Postępuj zgodnie z instrukcjami i celowo wchodź w interakcję ze sztuczną inteligencją, aby odkryć luki w zabezpieczeniach.
- Porównaj wyniki z innymi graczami i omów wyniki



<https://tensortrust.ai>



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Ćwiczenie praktyczne: Eksploracja bezpiecznych piaskownic AI



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Eksploracja bezpiecznych piaskownic AI

AI Dungeon

- Odwiedź stronę z grą do wstrzykiwania promptu: AI Dungeon
- Rozpocznij sesję
- Wejdź w interakcję ze sztuczną inteligencją, aby zobaczyć, jak radzi sobie z różnymi danymi wejściowymi.
- Omów mocne i słabe strony sztucznej inteligencji w bezpiecznym środowisku.

<https://play.aidungeon.com>



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Zastosowania w świecie rzeczywistym



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Zastosowania w świecie rzeczywistym

Rzeczywiste przykłady i ryzyko

- **Opieka zdrowotna:**

- **Ryzyko:** Narażenie danych pacjentów
- **Rozwiązanie:** Używaj silnej kontroli dostępu

- **Finanse:**

- **Ryzyko:** Oszustwa spowodowane błędami sztucznej inteligencji
- **Rozwiązanie:** uważne monitorowanie transakcji

- **Samochody autonomiczne:**

- **Ryzyko:** Nieprawidłowe reakcje na oszukańcze sygnały
- **Rozwiązanie:** Dokładne testowanie systemów AI



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)

Punkty do dyskusji

- Jak możemy wyprzedzić problemy związane z bezpieczeństwem sztucznej inteligencji?
- Które dziedziny są najbardziej zagrożone przez sztuczną inteligencję?
- Dlaczego ważne jest regularne sprawdzanie systemów AI?
- Co może się stać, jeśli bezpieczeństwo AI zostanie zignorowane?
- Które narzędzia lub praktyki są najlepsze dla bezpieczeństwa AI?
- Jak zrównoważyć nowe technologie AI z potrzebami bezpieczeństwa?



**Dofinansowane przez
Unię Europejską**

**WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)**

Dziękuję



**Dofinansowane przez
Unię Europejską**

WARSZTATY DLA MŁODYCH LUDZI "Technologies of tomorrow - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi" w ramach partnerskiego projektu współpracy na małą skalę w sektorze Młodzież programu Erasmus+ "Technologie jutra - zwalczanie dezinformacji i budowanie potencjału bezpieczeństwa w zakresie wykorzystania sztucznej inteligencji przez młodych ludzi".
(2023-1-ES02-KA210-YOU-000164824)