

Zrozumienie stronniczości algorytmicznej

Identyfikacja i ograniczanie
stronniczości w danych
i algorytmach sztucznej
inteligencji w celu
promowania
sprawiedliwości i równości.

Nr projektu: 2024-1-BG01-KA220-HED-000249185



Co-funded by
the European Union



Cele nauczania

W skrócie



- Zrozumienie, czym jest stronniczość algorytmiczna i skąd się bierze.
- Wskazanie różnych rodzajów stronniczości występujących w danych i systemach sztucznej inteligencji.
- Rozpoznanie zagrożeń związanych z wykorzystaniem sztucznej inteligencji w kontekście edukacyjnym.
- Analiza wpływu stronniczości na równość i inkluzywność.
- Stosowanie podstawowych strategii ograniczania stronniczości w praktyce dydaktycznej.



Najważniejsze koncepcje i definicje

*PODSTAWOWE POJĘCIA DOTYCZĄCE ROZPOZNAWANIA
I ELIMINOWANIA STRONNICZOŚCI W SYSTEMACH AI*

Stronniczość algorytmiczna

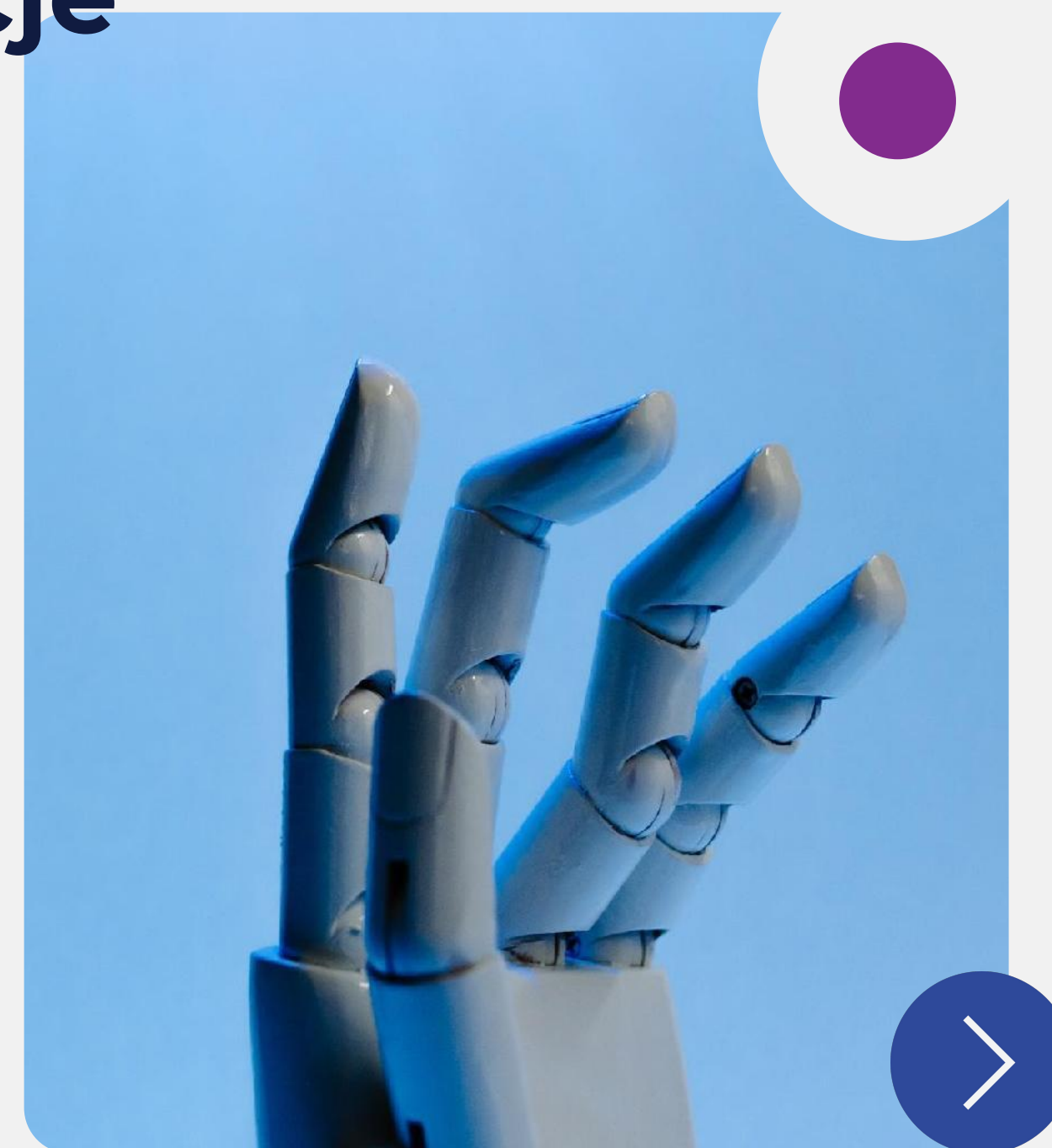
Stronniczość algorytmiczna występuje, gdy system AI systematycznie generuje niesprawiedliwe wyniki z powodu stronniczych danych, wyborów dokonanych na etapie projektowania lub ukrytych założeń.

Stronniczość reprezentacyjna

Stronniczość reprezentacyjna pojawia się, gdy niektóre grupy są niedostatecznie lub nadmiernie reprezentowane w danych szkoleniowych.

Stronniczość historyczna

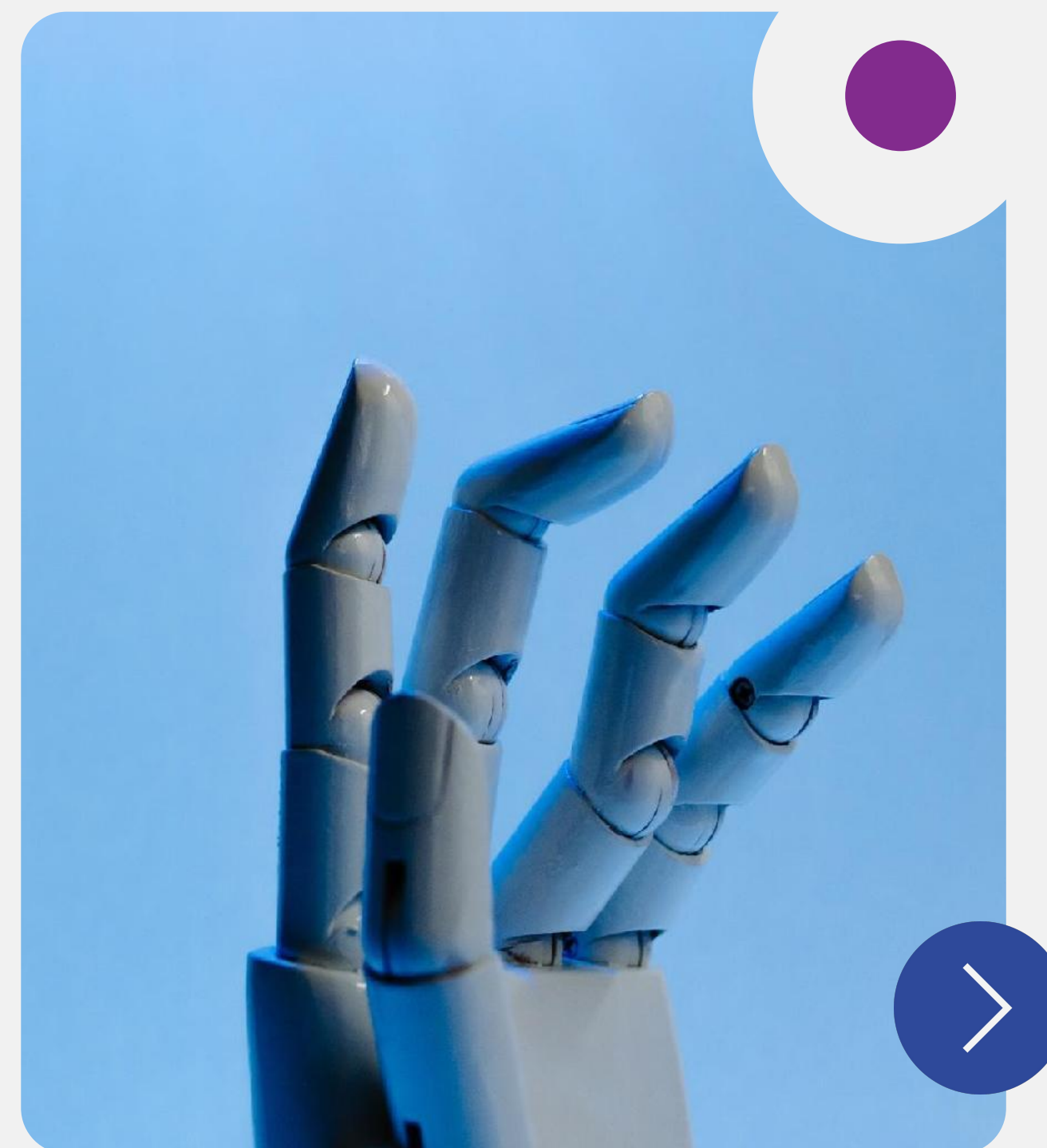
Stronniczość historyczna występuje, gdy zbiory danych odzwierciedlają dawne nierówności, które są odtwarzane lub wzmacniane przez systemy sztucznej inteligencji.



Czym jest stronniczość algorytmiczna?

Jak powstają zniekształcenia w danych i algorytmach AI.

- Do stronniczości algorytmicznej dochodzi, gdy system AI systematycznie generuje niesprawiedliwe lub nierówne wyniki z powodu danych, na których został wytrenowany, wyborów dokonanych na etapie projektowania lub sposobu, w jaki interpretuje informacje.
- Nie zawsze jest to zamierzone, ale może powodować utrwalanie lub pogłębianie istniejących nierówności.
- Stronniczość algorytmiczna to nie tylko kwestia techniczna, ale także odzwierciedlenie szerszych nierówności społecznych zakorzenionych w danych i praktykach instytucjonalnych.



Główne rodzaje stronniczości



Stronniczość historyczna

Gdy dane szkoleniowe odzwierciedlają dawne nierówności, które zostają zakodowane w systemie AI.

Stronniczość agregacji

Gdy jeden model stosowany jest do zróżnicowanych grup bez uwzględnienia istniejących różnic.

Stronniczość reprezentacyjna

Gdy pewne grupy są niedostatecznie lub nadmiernie reprezentowane w danych szkoleniowych.

Stronniczość pomiarowa

Gdy wykorzystywane dane nie oddają dokładnie tego, co mają oceniać.

Stronniczość projektowa lub wdrożeniowa

Wprowadzana przez ludzkie wybory przy definiowaniu zmiennych, kategorii lub kryteriów oceny.

Błąd confirmacji

Gdy programiści lub projektanci wybierają dane lub cechy wzmacniające przyjęte założenia.

Ryzyko związane ze stronniczością algorytmiczną

W jaki sposób dane mogą prowadzić do niesprawiedliwych wyników w systemach AI stosowanych w edukacji.



Stronniczość algorytmiczna często wynika z samych danych.

Gdy zbiory danych są niekompletne, niezrównoważone lub źle zorganizowane, systemy AI mogą generować zniekształcone lub niesprawiedliwe wyniki.

Główne zagrożenia związane z danymi:

- **Niedostateczna reprezentacja niektórych grup studentów** → doświadczenia i potrzeby niektórych studentów mogą nie być odpowiednio odzwierciedlone.
- **Niezrównoważone zbiory danych** → system może działać lepiej w przypadku grup większościowych niż mniejszościowych.
- **Wykorzystanie zmiennych zastępczych (wskaźników pośrednich)** → zmienne takie jak kod pocztowy lub typ szkoły mogą nieumyślnie odzwierciedlać różnice społeczno-ekonomiczne lub demograficzne.
- **Niskiej jakości lub niespójne dane** → niedokładne lub nieaktualne informacje mogą zniekształcać prognozy i rekomendacje.
- **Nadmierne poleganie na danych ilościowych** → złożona rzeczywistość edukacyjna może zostać sprowadzona do uproszczonych wskaźników.

Jeżeli dane są tendencyjne, nawet algorytm poparty solidnymi podstawami technicznymi może generować nierówne wyniki.



Wpływ stronniczości algorytmicznej na równość w edukacji



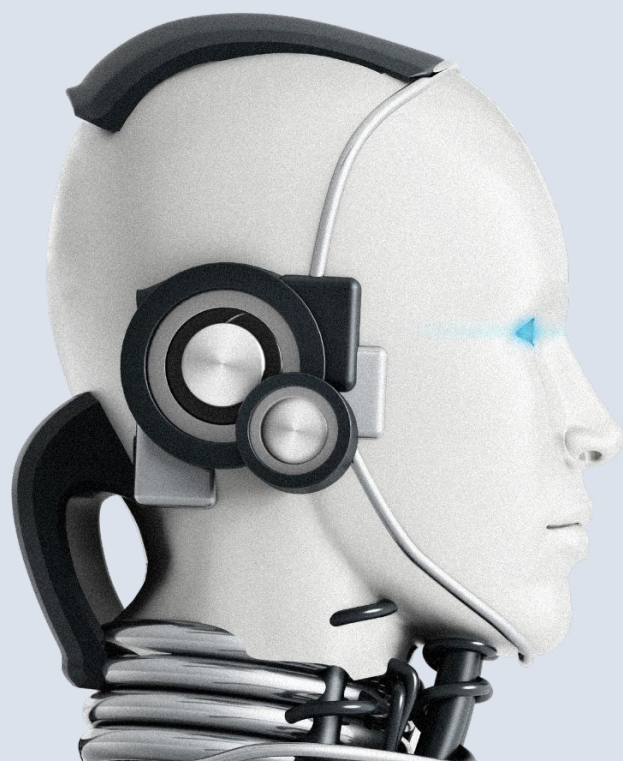
Jak systemy sztucznej inteligencji mogą wpływać na równość szans w szkolnictwie wyższym.

Gdy systemy AI zawierają elementy stronniczości, mogą w różny sposób oddziaływać na poszczególne grupy studentów, nawet bez wyraźnego zamiaru. Potencjalne skutki w kontekście edukacyjnym to:

- **Niesprawiedliwe lub nieproporcjonalne oceny.** Studenci o określonym profilu mogą konsekwentnie otrzymywać mniej korzystne oceny lub rekomendacje.
- **Wzmocnienie istniejących nierówności.** Historyczne wzorce mogą być odtwarzane w analizach predykcyjnych lub rekomendacjach dotyczących ścieżek akademickich.
- **Ograniczenie możliwości akademickich.** Systemy doradztwa lub monitorowania mogą wpływać na przyszłe wybory i postępy studentów.
- **Utrata zaufania do technologii edukacyjnych.** Postrzegana niesprawiedliwość może podważać zaufanie do narzędzi cyfrowych i praktyk instytucjonalnych.
- **Wpływ na inkluzywność i różnorodność.** Określone grupy mogą być systematycznie dyskryminowane.

System predykcyjny zaprojektowany w celu identyfikacji „studentów z grup ryzyka” może nieproporcjonalnie często wskazywać studentów ze środowisk znajdujących się w niekorzystnej sytuacji, jeśli historyczne dane dotyczące porzucania nauki odzwierciedlają istniejące nierówności strukturalne.





Praktyczne działania mające na celu ograniczanie stronniczości algorytmicznej i promowanie równości w edukacyjnej sztucznej inteligencji

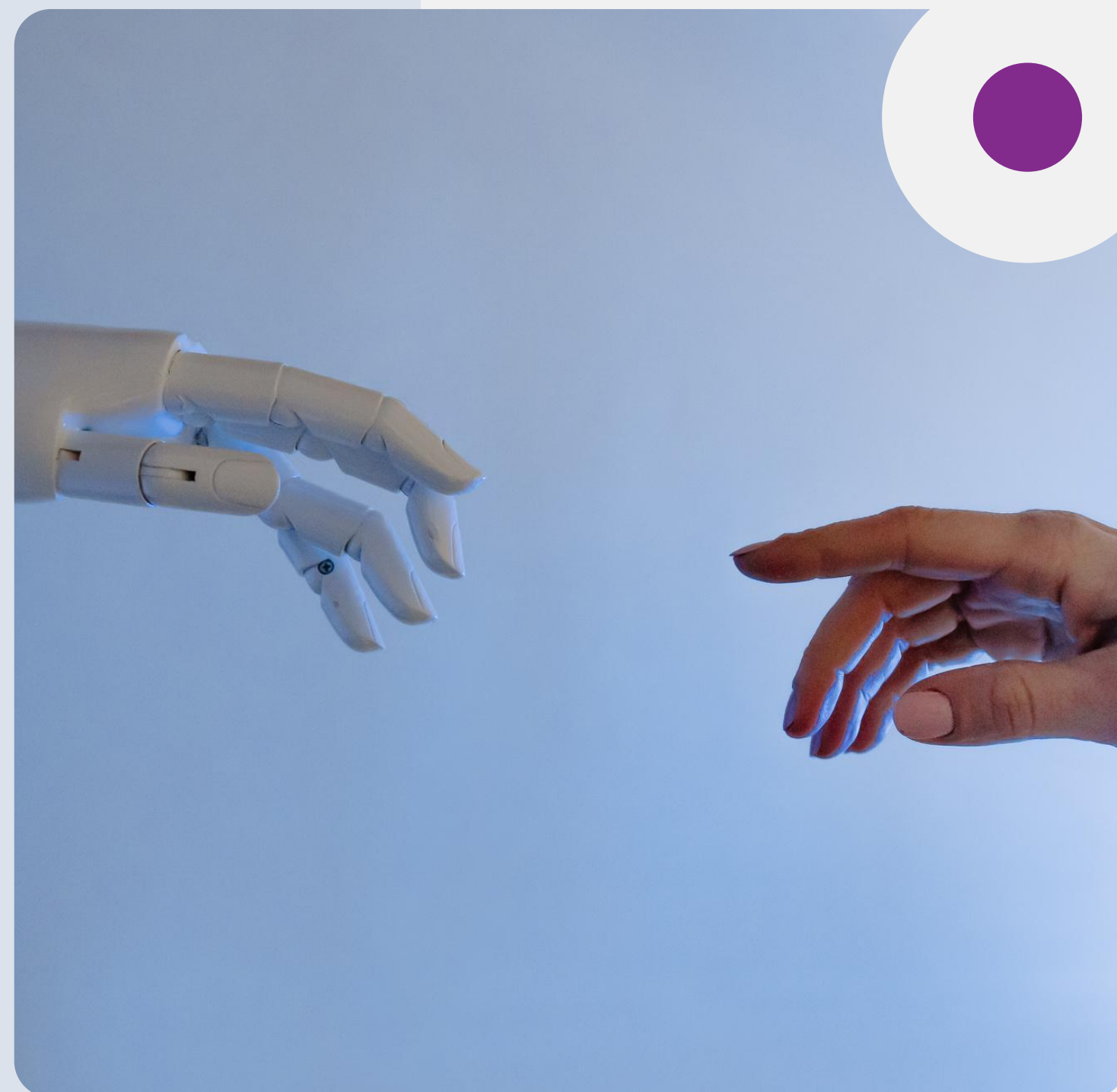
Nr projektu: 2024-1-BG01-KA220-HED-000249185

Chociaż nie zawsze da się całkowicie wyeliminować stronniczość, można ją identyfikować i ograniczać poprzez świadome decyzje podejmowane w praktyce dydaktycznej.

- Sprawdzaj, czy wykorzystywane dane odpowiednio odzwierciedlają profile wszystkich studentów.
- Zapewnij aktywny nadzór ze strony ludzi i weryfikuj rekomendacje generowane przez sztuczną inteligencję przed podjęciem decyzji akademickich.
- Promuj przejrzystość, wyjaśniając studentom, w jaki sposób wykorzystywane są narzędzia oparte na sztucznej inteligencji i jakie są ich ograniczenia.
- Zwracaj uwagę, czy określone grupy są systematycznie narażone na niekorzystne skutki.
- Weryfikuj reprezentatywność danych, biorąc pod uwagę, czy system został wytrenowany na zróżnicowanych zbiorach danych, czy niektóre grupy mogą być niedostatecznie reprezentowane.
- Sprawdzaj, czy system systematycznie generuje różne wyniki w zależności od płci, pochodzenia, języka lub innych istotnych czynników.
- Słuchaj opinii studentów, by ustalić, czy któraś z grup odczuwa nierówne traktowanie.



Co-funded by
the European Union



Czy potrafisz rozpoznać stronniczość?

Wykorzystaj zdobytą wiedzę

Stronniczość algorytmiczna nie zawsze jest oczywista. Często przejawia się w subtelny sposób w narzędziach edukacyjnych i systemach decyzyjnych. W ramach tego zadania przeanalizujesz rzeczywiste scenariusze edukacyjne i zidentyfikujesz rodzaj występującej w nich stronniczości.

Wykorzystaj swoją wiedzę w praktyce i sprawdź swoje umiejętności krytycznej oceny.

Rozpocznij ćwiczenie



Źródła

- Barocas, S., Hardt, M. i Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*.
<https://fairmlbook.org>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. i Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.
<https://doi.org/10.1145/3457607>
- UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence (Zalecenia dotyczące etyki sztucznej inteligencji)*.
<https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- Komisja Europejska. (2022). *Ethical guidelines on the use of artificial intelligence and data in teaching and learning for educators* (Wytyczne etyczne dotyczące wykorzystania sztucznej inteligencji i danych w nauczaniu i uczeniu się, przeznaczone dla nauczycieli).
<https://education.ec.europa.eu/news/ethical-guidelines-on-the-use-of-artificial-intelligence-and-data-in-teaching-and-learning-for-educators>



Promocja równości w edukacji wymaga nie tylko dobrych intencji, ale także krytycznej świadomości tego, jak technologia wpływa na dostęp do możliwości.

Konsorcjum:



<https://trustai.unibit.bg/>



Projekt Trust AI

Wyłączenie odpowiedzialności: Finansowane przez Unię Europejską. Wyrażone poglądy i opinie stanowią wyłącznie poglądy autora (autorów) i niekoniecznie odzwierciedlają poglądy Unii Europejskiej lub Narodowej Agencji (NA). Ani Unia Europejska, ani NA nie ponoszą za nie odpowiedzialności.

Nr projektu: 2024-1-BG01-KA220-HED-000249185



Co-funded by
the European Union

