

ZAJĘCIA 3

Kwestie związane z etyką AI, uprzedzeniami, prywatnością, środowiskiem i kontrolą nad technologią

Ważne pytania i odpowiedzialność związane z AI,
które biblioteki powinny rozumieć i świadomie podejmować.

© 60–90 min | Wszystkie osoby pracujące w bibliotekach

Dlaczego te pytania są pilne?

Rozwój AI postępuje szybciej niż refleksja etyczna. Biblioteki powinny wyprzedzać te zmiany, a nie jedynie próbować za nimi nadążyć.



Problemy etyczne już występują

Systemy rozpoznawania twarzy błędnie identyfikowały niewinne osoby. Algorytmy rekrutacyjne systematycznie dyskryminowały kobiety. To nie są hipotetyczne scenariusze. Takie sytuacje dzieją się już teraz.



Uprzedzenia nie są przypadkowym błędem systemu

Wynikają z danych, na których sztuczna inteligencja się uczy. Jeśli dane odzwierciedlają nierówności społeczne, AI również będzie je odtwarzać. Bez świadomej interwencji problem nie zniknie sam.



Dane użytkowników są zagrożone

Każdy system AI analizujący zapytania użytkowników, historię wyszukiwania lub czytania rodzi poważne pytania dotyczące prywatności i poufności danych.



AI ma realny wpływ na środowisko

Wytrenowanie jednego dużego modelu AI może generować emisję CO₂ porównywalną z całym cyklem życia pięciu samochodów. Biblioteki deklarujące troskę o zrównoważony rozwój powinny uwzględniać również ten aspekt.



Kto kontroluje AI, kontroluje wiedzę

Niewielka liczba globalnych firm kontroluje dziś systemy AI wpływające na sposób wyszukiwania, filtrowania i prezentowania informacji. To nie tylko kwestia technologii, ale także kontroli nad wiedzą i dostępem do informacji.



Etyka AI

Co jest właściwe? Co może być szkodliwe? Kto o tym decyduje?

Etyka nie ogranicza AI, lecz pomaga odpowiedzialnie wykorzystywać tę technologię.

Globalne ramy etyczne AI

Powstały już międzynarodowe wytyczne dotyczące odpowiedzialnego wykorzystania AI. Bibliotekarze powinni je rozumieć i wykorzystywać w swojej pracy.

UNESCO (2021)

Godność człowieka i inkluzywność

Pierwszy globalny dokument normatywny dotyczący etyki AI. Podkreśla znaczenie praw człowieka, równości płci, zrównoważonego rozwoju i bezpiecznych społeczeństw. Wyraźnie odnosi się także do wykorzystania AI w edukacji i dostępie do informacji.

EU AI Act (2024)

Regulacje oparte na poziomie ryzyka

Systemy AI są klasyfikowane według poziomu ryzyka: niedopuszczalne, wysokiego ryzyka, ograniczonego ryzyka i minimalnego ryzyka. Biblioteki wykorzystujące AI np. do profilowania użytkowników lub wspierania decyzji kadrowych mogą zostać zakwalifikowane do kategorii systemów wysokiego ryzyka.

IFLA (2020)

Dostęp, prywatność i zaufanie

Stanowisko Międzynarodowej Federacji Stowarzyszeń Bibliotekarskich podkreśla znaczenie otwartego dostępu do informacji, ochrony prywatności, wolności od nadzoru oraz prawa do korzystania z informacji bez algorytmicznego pośrednictwa i manipulacji.

OECD (2019)

Inkluzywny rozwój i odpowiedzialność

Pięć głównych zasad obejmuje: inkluzywny rozwój, wartości skoncentrowane na człowieku, przejrzystość, bezpieczeństwo, odporność systemów oraz odpowiedzialność. Zasady te zostały przyjęte przez państwa G20 i są dziś obecne w większości krajowych strategii dotyczących AI.

W jaki sposób AI powoduje realne szkody?

Zrozumienie różnych sposobów, w jakie systemy AI mogą powodować realne szkody.

Typ szkody	Przykład z życia	Znaczenie dla bibliotek
Szkody związane z dostępem do zasobów i usług	AI odmawia komuś kredytu, pracy lub świadczeń mieszkaniowych na podstawie stronniczych przewidywań.	System AI może ograniczać dostęp do usług bibliotecznych lub oznaczać konta użytkowników na podstawie wzorców demograficznych.
Szkody reprezentacyjne	AI kojarzy określone grupy z negatywnymi cechami, np. poprzez rasowo stronnicze obrazy „przestępców”.	Systemy AI wspierające wyszukiwanie w katalogach mogą słabiej promować literaturę z krajów Globalnego Południa.
Szkody dla jakości usług	Rozpoznawanie głosu działa gorzej w przypadku kobiet i osób niebędących rodzimymi użytkownikami języka.	Narzędzia AI mogą działać lepiej dla osób anglojęzycznych niż dla innych użytkowników bibliotek.
Szkody wynikające z manipulacji i nadużyć	Deepfake’i mogą być wykorzystywane do nękania, manipulacji lub oczerniania ludzi.	Użytkownicy mogą stać się celem dezinformacji generowanej przez AI na podstawie danych powiązanych z aktywnością biblioteczną.
Szkody związane z dezinformacją i utratą zaufania do wiedzy	AI potrafi przedstawiać nieprawdziwe informacje w bardzo przekonujący sposób, co utrudnia odróżnianie faktów od błędów.	Takie treści mogą później trafiać do badań naukowych i debaty publicznej.



Stronniczość algorytmiczna

Stronniczość nie jest przypadkowym błędem systemu.
Często jest skutkiem danych, na których system się uczy.

AI uczy się na świecie takim, jaki jest, razem z istniejącymi
w nim nierównościami i uprzedzeniami.

Źródła stronniczości w systemach AI

Uprzedzenia mogą pojawiać się na każdym etapie tworzenia systemów AI, nie tylko na etapie danych treningowych.

Zbieranie danych

Brak odpowiedniej reprezentacji grup mniejszościowych, historyczne uprzedzenia zapisane w istniejących danych, dane treningowe pobierane z internetu skupione głównie na perspektywach bogatszych, anglojęzycznych społeczeństw.

Oznaczanie danych

Osoby opisujące dane wnoszą własne założenia kulturowe. Kategorie tworzone podczas oznaczania danych częściej odzwierciedlają punkt widzenia osób wykonujących tę pracę niż społeczności, których dane dotyczą.

Projektowanie modelu

AI może działać świetnie dla większości osób, ale jednocześnie zawodzić wobec mniej reprezentowanych grup. Już sama architektura modelu zawsze opiera się na określonych założeniach.

Kontekst wdrożenia

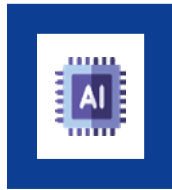
System wytrenowany w jednym środowisku może działać błędnie w innym. Przykładowo model AI wytrenowany na danych pacjentów z USA może nie sprawdzać się równie dobrze w krajach Afryki Subsaharyjskiej.

Pętle wzmacniające

Jeśli stronnicze wyniki AI stają się częścią nowych danych treningowych, istniejące uprzedzenia mogą się z czasem się pogłębić i utrwaląć (np. stronniczy system rekrutacyjny → stronnicze dane o zatrudnieniu → kolejne modele AI trenowane na tych danych).

Stronniczości algorytmiczna

« Grupa golfistów »



Stronniczości algorytmiczna



Stronniczości algorytmiczna

≡ Przetłumacz

English ▼

Turkish ▼

He is a doctor



0 bir doktor



Stronniczości algorytmiczna

Garbage in, garbage out (GIGO) (pol. śmieci na wejściu – śmieci na wyjściu)

Jeśli AI uczy się na złych danych, będzie generować złe wyniki.



Stronniczości algorytmiczna

Holenderski skandal służy jako ostrzeżenie dla Europy przed ryzykiem stosowania algorytmów

Holenderski urząd skarbowy zrujnował życie tysiącom ludzi po użyciu algorytmu do wykrywania podejrzeń o oszustwa zasiłkowe – a krytycy twierdzą, że niewiele może powstrzymać to przed ponownym powtórzeniem się

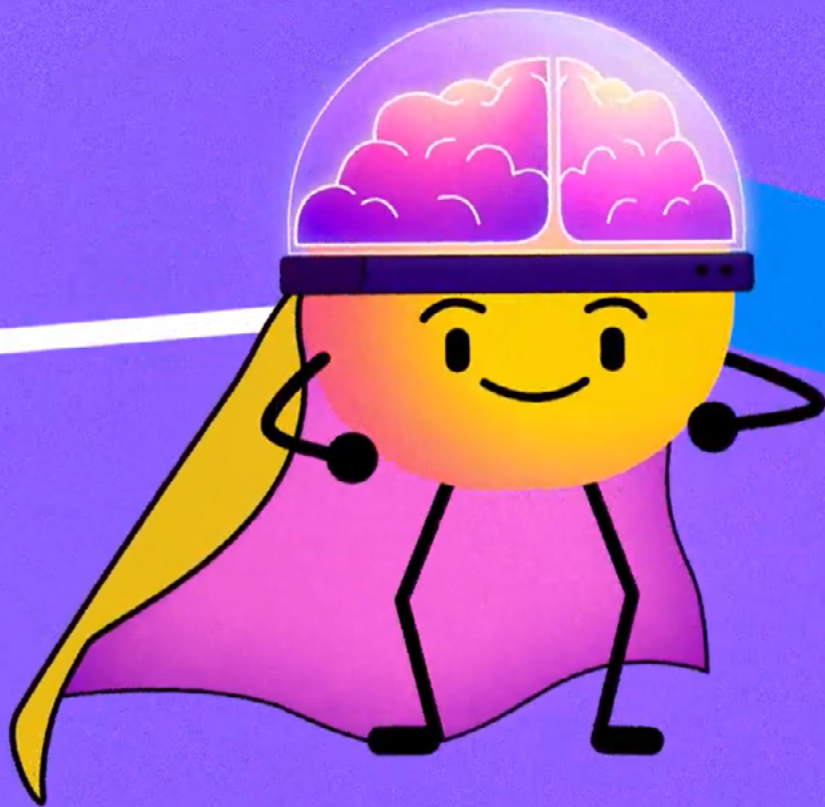
**SCORING BENEFICJENTÓW OPIEKI
SPOŁECZNEJ: NIEGODZIWOŚĆ ALGORYTMU
CAF JEST JUŻ NIEZAPRZECZALNA**

27 listopada 2023

Stronniczości algorytmiczna

- System rekrutacyjny Amazona dyskryminował kobiety, ponieważ uczył się na historycznych danych z branży technologicznej zdominowanej przez mężczyzn.
- Badania dotyczące systemów używanych przez organy ścigania wykazały nawet o 34% wyższy poziom błędów przy rozpoznawaniu twarzy osób o ciemniejszym kolorze skóry.
- Niektóre algorytmy medyczne gorzej oceniały potrzeby zdrowotne pacjentów czarnoskórych.
- Reklamy mieszkań i ofert pracy mogą być kierowane do określonych grup demograficznych na podstawie algorytmów.

Stronniczości algorytmiczna





Prywatność danych i nadzór

„Jeśli nie płacisz za produkt, sam jesteś produktem.”

Dane użytkowników bibliotek nie są zasobem do wykorzystania.
Są prawem, które należy chronić.

Jak systemy AI gromadzą i wykorzystują dane

Przepływy danych bywają niewidoczne, ale ich konsekwencje już nie.

Łańcuch przepływu danych

- 1 Użytkownik wysyła zapytanie przez chat lub wyszukiwarkę opartą na AI.
- 2 Zapytanie trafia do zewnętrznego dostawcy AI (często zlokalizowanego w innym kraju).
- 3 Dostawca może zapisywać zapytania w celu dalszego trenowania modeli.
- 4 Dane o sposobie korzystania z systemu są agregowane i potencjalnie sprzedawane.
- 5 Powstają profile łączące historię wypożyczeń, wyszukiwań i zapytań użytkownika.
- 6 W określonych systemach prawnych takie profile mogą być dostępne dla organów władzy.

Istniejące zabezpieczenia

- ▶ RODO (UE): prawo dostępu do danych, usunięcia danych i minimalizacji przetwarzania
- ▶ W wielu krajach przepisy dotyczące poufności danych bibliotecznych obejmują również interakcje cyfrowe
- ▶ Ramy prywatności IFLA dla systemów bibliotecznych

Kluczowe ryzyka

- ▶ Przekazywanie danych do innych krajów może nie zapewniać takiego samego poziomu ochrony prywatności
- ▶ Narzędzia AI mogą przyznawać dostawcom szerokie prawa do wykorzystywania zapytań użytkowników
- ▶ Zapytania użytkowników mogą ujawniać wrażliwe dane dotyczące zdrowia, sytuacji prawnej lub poglądów politycznych

Jak systemy AI gromadzą i wykorzystują dane

Wykorzystanie AI nie zwalnia bibliotek z obowiązku ochrony prywatności użytkowników.

PODSTAWOWA ZASADA: Odpowiedzialność biblioteki za prywatność użytkowników nie kończy się tam, gdzie zaczyna się AI. Każdy system AI przetwarzający dane użytkowników powinien spełniać takie same standardy ochrony jak każdy inny system biblioteczny.

1

Ocena wpływu na prywatność danych: Przed wdrożeniem narzędzia AI mającego kontakt z danymi użytkowników należy przeprowadzić udokumentowaną ocenę skutków dla ochrony danych (DPIA). Należy ocenić, jakie dane są zbierane, gdzie trafiają oraz jakie ryzyka mogą się z tym wiązać.

2

Minimalizacja danych: Narzędzia AI powinny być skonfigurowane tak, aby zbierały wyłącznie minimalną ilość niezbędnych danych. W miarę możliwości należy unikać umieszczania danych umożliwiających identyfikację osób w promptach i zapytaniach kierowanych do AI.

3

Analiza polityk prywatności dostawców: Przed wdrożeniem narzędzia AI należy dokładnie przeczytać politykę prywatności i warunki korzystania z usług. Kluczowe pytania: Czy dostawca przechowuje zapytania użytkowników? Jak długo przechowywane są dane? Czy dane mogą być wykorzystywane do dalszego trenowania modeli AI?

4

Informowanie użytkowników: Jeśli sztuczna inteligencja jest wykorzystywana w usługach dla użytkowników, biblioteka powinna jasno o tym informować.

5

Szkolenie pracowników z zakresu prywatności i AI: Każda osoba korzystająca z narzędzi AI powinna wiedzieć, jakie dane można wprowadzać do systemu, a jakich danych nie wolno udostępniać. Warto przygotować listę dozwolonych oraz zabronionych typów danych.



Zadanie

Śledzenie plików cookies i ustawienia prywatności

[Zobacz instrukcję dla prowadzącego](#)



Wpływ na środowisko

AI ma realny ślad środowiskowy, który biblioteki deklarujące troskę o zrównoważony rozwój powinny brać pod uwagę.

Koszt energetyczny sztucznej inteligencji kryje się w serwerowniach i centrach danych.

Wpływ na środowisko

Współczesne systemy AI działają na ogromną skalę i wymagają bardzo dużych zasobów energii oraz infrastruktury.

626 000

kg CO₂

Szacowana emisja związana z wytrenowaniem jednego dużego modelu językowego od podstaw (badanie MIT)

10×

więcej energii

Jedno zapytanie do ChatGPT może zużywać około 10 razy więcej energii niż tandardowe wyszukiwanie w Google

6,4 mln

m³ wody

Szacowane zużycie wody przez centra danych firmy Microsoft w 2022 roku do chłodzenia systemów AI

2–4%

globalnych emisji CO₂

Obecny udział sektora ICT w światowych emisjach dwutlenku węgla szybko rosnący wraz z rozwojem AI



Władza i kontrola nad AI

Kto tworzy systemy AI? Kto kontroluje sztuczną inteligencję?
Czyje wartości odzwierciedla?

Kontrola nad infrastrukturą AI oznacza również kontrolę
nad przyszłością wiedzy i informacji.

Problem koncentracji rynku

Pięć firm: OpenAI (Microsoft), Google DeepMind, Anthropic (Amazon), Meta AI i xAI, kontroluje dziś najbardziej zaawansowane modele AI. Ich decyzje dotyczące danych treningowych, dostępu do modeli, cen, moderowania treści i zasad działania systemów wpływają na to, jakie informacje miliardy ludzi mogą wyszukiwać, oglądać i wykorzystywać.

Koncentracja mocy obliczeniowej	Trenowanie dużych modeli AI wymaga ogromnych zasobów obliczeniowych należących do niewielkiej liczby dostawców chmurowych (AWS, Azure, Google Cloud).
Koncentracja danych	Większość danych wykorzystywanych do trenowania modeli językowych pochodzi głównie z bogatszych, anglojęzycznych krajów.
Koncentracja pracy	Moderowanie treści i oznaczanie danych dla AI jest często wykonywane przez słabo opłacanych pracowników z krajów Globalnego Południa, nierzadko bez odpowiedniej ochrony pracowniczej.
Koncentracja wartości	Decyzje dotyczące tego, co AI może lub nie może powiedzieć, pokazać czy rekomendować, są podejmowane przez niewielką liczbę firm i zespołów działających w kilku krajach.
Nierówności regulacyjne	Regulacje dotyczące AI rozwijają się najszybciej w Unii Europejskiej. W wielu innych regionach świata nadzór nad systemami AI pozostaje ograniczony, mimo że wpływają one na życie milionów ludzi

Suwerenność AI: znaczenie dla bibliotek

Suwerenność kolekcji: Gdy AI zaczyna filtrować, porządkować lub rekomendować zbiory biblioteczne, część decyzji o tym, co użytkownicy widzą, przechodzi z bibliotekarzy do algorytmów tworzonych głównie dla celów komercyjnych.

→ *Zapewnij nadzór pracowników nad tym, jak AI wpływa na widoczność zbiorów.*

Regularnie analizuj sposób tworzenia rankingów.

Suwerenność wyszukiwania: AI wpływa na to, jakie informacje użytkownicy znajdują i widzą w pierwszej kolejności, nawet jeśli same zasoby się nie zmieniają. Jeśli użytkownik korzysta wyłącznie z odpowiedzi AI, jego dostęp do informacji jest częściowo kształtowany przez sposób działania modelu, a nie tylko przez jego rzeczywiste potrzeby.

→ *Oferuj także ścieżki wyszukiwania niewykorzystujące AI. Ucz użytkowników korzystania z haseł przedmiotowych.*

Suwerenność wiedzy: Systemy AI trenowane głównie na danych z Globalnej Północy mogą przedstawiać jedną perspektywę kulturową jako uniwersalną.

Biblioteki pracujące z różnorodnymi społecznościami

→ *Wzmacniaj mniej reprezentowane tradycje wiedzy. Unikaj AI spłaszczającej różnorodność kulturową.*

Suwerenność prawna: Dane przetwarzane przez dostawców AI mogą podlegać zagranicznym systemom nadzoru.

Biblioteki powinny sprawdzać, czy dane przetwarzane przez zagraniczne systemy AI są chronione zgodnie z krajowymi przepisami o prywatności.

→ *Przed wdrożeniem AI warto konsultować kwestie przechowywania i przepływu danych z ekspertami prawnymi.*

Ćwiczenie grupowe: audyt etyczny

Grupy 4–6 osób | 25 minut | Wybierzcie jedno narzędzie AI używane lub rozważane do wdrożenia w Waszej bibliotece.



Etyka

Czy narzędzie posiada opublikowane zasady etyczne?
Kto je opracował?
Jakie problemy i ryzyka związane z tym narzędziem zostały już opisane?



Stronniczość

Czy narzędzie było niezależnie testowane pod kątem uprzedzeń wobec różnych grup społecznych?
Jakie języki i społeczności są słabiej reprezentowane w danych treningowych?



Prywatność

Jakie dane są zbierane i jak długo są przechowywane?
Czy narzędzie jest zgodne z RODO? Czy dostępna jest odpowiednia dokumentacja?



Środowisko

Czy dostawca publikuje dane dotyczące zużycia energii lub emisji CO₂?
Czy centra danych są zasilane energią odnawialną?



Suwerenność

W jakich krajach przetwarzane są dane?
Czy polityka moderowania treści dostawcy jest zgodna z wartościami i prawami naszej społeczności?