Uprzedzenia i sprawiedliwość w generatywnej sztucznej inteligencji

Streszczenie

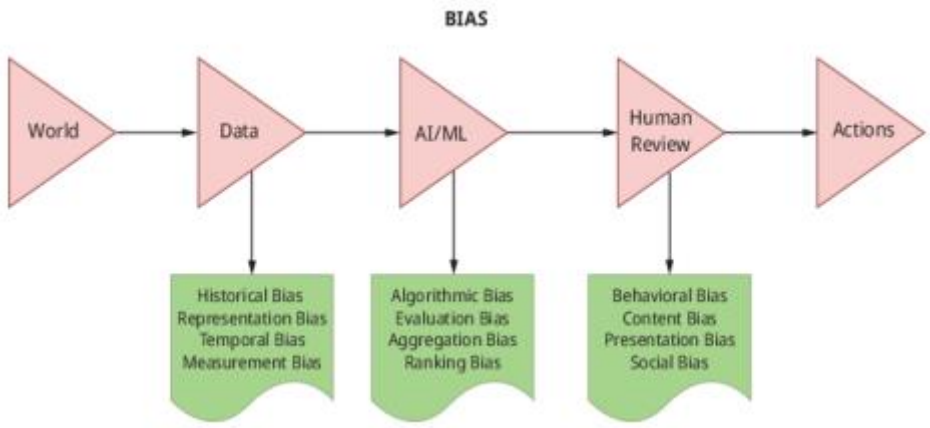
Omówiono zagadnienia uprzedzeń i sprawiedliwości w generatywnej sztucznej inteligencji, przedstawiając kompleksowy obraz głębokiego związku między sztuczną inteligencją a ludzkimi uprzedzeniami oraz ich wpływu na postęp i wdrażanie generatywnych prototypów. Jest to dogłębna analiza uprzedzeń w danych treningowych, algorytmicznym podejmowaniu decyzji i niezamierzonych konsekwencjach zastosowań sztucznej inteligencji, z uwzględnieniem przykładów z życia wziętych. Zbadanie głębokich komplikacji związanych z uprzedzeniami w generatywnych systemach sztucznej inteligencji oraz możliwych rozwiązań będzie dominować w tym rozdziale. Postaramy się odnieść te dyskusje do kwestii etycznych i odpowiedzialności związanych z ich generowaniem i użytkowaniem. Dałoby nam to bardzo jasny fundament do dyskusji o tym, jak należy równoważyć innowacje i kwestie etyczne w dziedzinie sztucznej inteligencji. Ten humanistyczny punkt widzenia ustąpi miejsca potrzebie skupienia się na szerokich implikacjach społecznych dotyczących rozwoju sztucznej inteligencji.

Słowa kluczowe: Uprzedzenia, generatywna sztuczna inteligencja, rozważania etyczne, ograniczanie uprzedzeń, ograniczanie sprawiedliwości.

Wprowadzenie

Błąd systematyczny

Generatywne modele sztucznej inteligencji, które uwzględniają błąd systematyczny, o którym mowa w tym miejscu, to te, które wynikają z poważnych błędów modelu o charakterze systematycznym, skutkujących niesprawiedliwym lub dyskryminującym wynikiem. Takie błędy systematyczne mogą wynikać z procesów optymalizacji, projektu architektury lub danych szkoleniowych. Formy, w jakich błędy systematyczne mogą się przejawiać w wygenerowanym modelu sztucznej inteligencji, to negatywne narracje, utrwalanie stereotypów lub nierównowaga w populacji. Wykorzystanie systemów sztucznej inteligencji na dużą skalę doprowadziło do kontrowersyjnego pytania o uczciwość i błąd systematyczny w sztucznej inteligencji, a także do gorących debat na temat źródeł, skutków i metod łagodzenia. Obecność błędów systematycznych w wykorzystaniu sztucznej inteligencji dopiero niedawno stała się przedmiotem debaty i dyskusji. Identyfikacja i łagodzenie błędów systematycznych w danych to nie tylko problemy matematyczne, ale także kwestia społecznej złożoności uczciwości, zmieniającej się w zależności od sytuacji i odzwierciedlającej wartości, etykę i zobowiązania prawne. Istnieją jasne metody dbania o sprawiedliwość sztucznej inteligencji; kroki podejmowane w celu ograniczenia stronniczości zależą od tego, czego chcesz uniknąć w swoim modelu. Rysunek przedstawia przegląd stronniczości w sztucznej inteligencji. Reprezentuje on różne rodzaje stronniczości w różnych środowiskach.



Sprawiedliwość

Sprawiedliwość i stronniczość w sztucznej inteligencji (AI) nie mają uniwersalnej definicji, ale generalnie odnoszą się do braku faworyzowania lub uprzedzeń wobec jednostek lub grup ze względu na ich atrybuty. Pozwólcie, że zilustruję to przykładem, jak algorytm uczenia maszynowego może napotkać problemy ze sprawiedliwością. Termin „sprawiedliwość" jest jednym z najczęściej poruszanych tematów w AI i ML. Jest on elementem większości odpowiedzialnych i etycznych koncepcji AI. Co jednak dokładnie oznacza ten termin i co oznacza, że system uczenia maszynowego jest sprawiedliwy? Ten krótki rozdział omawia „sprawiedliwość" w ogólności, zanim przejdziemy do domyślnego podejścia do sprawiedliwości w uczeniu maszynowym i związanych z nim trudności. Sprawiedliwość w sztucznej inteligencji spotkała się z dużym zainteresowaniem zarówno ze strony środowiska akademickiego, jak i biznesowego. Sprawiedliwość w sztucznej inteligencji oznacza, że istnieją systemy wolne od uprzedzeń i dyskryminacji, co stanowi bardzo poważny problem w tego typu systemach. W niniejszym artykule omówiono kilka form sprawiedliwości: kontrfaktyczną, indywidualną i grupową. Są one ze sobą powiązane, ale sprawiedliwość i uprzedzenia to zniuansowane rozróżnienia. Innymi słowy, stronniczość może wystąpić nieumyślnie, ale uczciwość nigdy nie może być niczym innym niż zamierzonym i celowym działaniem. Sprawiedliwość w AI, w kontekście innych interesariuszy, musi zostać starannie przemyślana. Te przykłady uczciwości w rzeczywistych zastosowaniach AI pokazują potencjalne korzyści . Niniejsza część omawia uprzedzenia danych, algorytmów i użytkowników, zapewniając szczegółowy przegląd przyczyn i skutków uprzedzeń w AI, a także podkreślając niektóre konsekwencje etyczne. Omawia on stan wiedzy w zakresie badań nad strategiami łagodzenia, odnosząc się do niektórych z ich ograniczeń, trudności i wartości współpracy multidyscyplinarnej. Uczciwość i stronniczość to kluczowe zagadnienia w AI, formułowane przez badaczy, prawodawców i społeczność akademicką. Niniejsze badanie przeglądowe analizuje złożone i rozbieżne kwestie uprzedzeń i uczciwości w AI, związane z przyczynami uprzedzeń, ich skutkami i możliwymi środkami łagodzenia uprzedzeń. W rozdziale tym starano się wyjaśnić przyczyny, skutki i metody łagodzenia skutków uczciwości i stronniczości w dziedzinie sztucznej inteligencji, a tym samym wspierać ciągłe wysiłki mające na celu tworzenie bardziej etycznych i odpowiedzialnych systemów sztucznej inteligencji.

Błąd systematyczny: źródła, wpływ i strategie łagodzenia

Źródła błędu systematycznego

Jednym z niesystematycznych błędów w procesie podejmowania decyzji przez system jest błąd systematyczny. Źródła błędu systematycznego w sztucznej inteligencji są liczne i obejmują ludzką interpretację, projektowanie algorytmów i gromadzenie danych. Przykładem systemu sztucznej inteligencji generującego błąd systematyczny jest model uczenia maszynowego, który replikuje błąd

systematyczny występujący w danych treningowych. Taki system zapewnia zachowanie sprawiedliwości i równości w tych systemach dla każdej osoby. W kolejnych rozdziałach zostaną przedstawione dalsze badania nad przyczynami, skutkami i sposobami łagodzenia błędu systematycznego w sztucznej inteligencji. Błąd systematyczny – błąd systemowy powodujący niesprawiedliwe wyniki w procesie podejmowania decyzji – powstaje na skutek szeregu czynników, od interpretacji ludzkiej, przez projektowanie, po gromadzenie danych. Wśród źródeł błędu systematycznego, czyli niesprawiedliwości, należy zauważyć, że jedną z form systemu sztucznej inteligencji generującego błąd systematyczny jest model uczenia maszynowego, który replikuje błąd systematyczny występujący w danych treningowych. W związku z tym można zapewnić, że systemy te będą sprawiedliwe i równe dla każdej osoby, biorąc pod uwagę rozpoznanie i eliminację uprzedzeń w sztucznej inteligencji. Uprzedzenia w sztucznej inteligencji (AI) w algorytmach uczenia maszynowego wynikają z błędnych założeń i zniekształconych zbiorów danych, będąc jednocześnie zgodnymi z istniejącymi ludzkimi uprzedzeniami, a czasami je wzmacniając. Na przykład algorytmy rozpoznawania twarzy mogą zawierać uprzedzenia ze względu na płeć, jeśli są trenowane na danych dotyczących twarzy, na których w dużej mierze znajdują się mężczyźni, a prawie wcale kobiety. W podobny sposób mogłoby to odzwierciedlać złożoną dynamikę czynników płci i cech, które z natury różnią kobiety od mężczyzn. W procesach rekrutacji, algorytmy mogą dyskryminować osoby o ciemnej karnacji lub obcojęzycznych nazwiskach, nie rozumiejąc ich kwalifikacji zawodowych. To uprzedzenie nie tylko przywraca nierówności społeczne, ale także utrudnia równe szanse grupom mniejszościowym, tworząc trudny do przerwania cykl dyskryminacji. Co więcej, nieświadome uprzedzenia jedynie pogarszają wyzwanie związane z sprawiedliwą sztuczną inteligencją, która jest subtelna i trudna do wykrycia. Te niechciane uprzedzenia, w procesie rozwoju i użytkowania sztucznej inteligencji, mogą niepostrzeżenie utrwalać nierówności systemowe bez wyraźnej wiedzy i intencji. Radzenie sobie z uprzedzeniami w sztucznej inteligencji będzie wymagało holistycznego podejścia, obejmującego staranne wstępne przetwarzanie danych, transparentność stosowanych algorytmów oraz okresowe oceny w celu radzenia sobie z negatywnym wpływem zarówno świadomych, jak i nieświadomych uprzedzeń na systemy uczenia maszynowego, zapewniając w ten sposób sprawiedliwe i równe wyniki dla wszystkich. Dalsze badania dotyczące przyczyn, skutków i sposobów łagodzenia uprzedzeń w sztucznej inteligencji będą dostępne w kolejnych sekcjach.

Błąd danych

Błąd danych to systematyczny błąd, który pojawia się w modelach sztucznej inteligencji (AI) w wyniku wykorzystania niereprezentatywnych lub skośnych danych do trenowania. Innymi słowy, jest to stan, w którym dane treningowe nie odpowiadają rzeczywistemu scenariuszowi lub populacji, dla której model AI jest przeznaczony . Ta rozbieżność między danymi użytymi do trenowania a rzeczywistymi scenariuszami prowadzi do fałszywych prognoz lub niesprawiedliwych decyzji lub rekomendacji wydawanych przez system AI. Jest to zatem niezbędne do tworzenia systemów AI, które działają bezbłędnie, z wiarygodnością i rzetelnością dla szerokiego grona osób i kontekstów. W tym kontekście błąd danych może stać się bardzo szeroką kategorią, poprzez którą mogą potencjalnie wpływać na wydajność i rzetelność AI. Na przykład: Błąd doboru próby odnosi się do zjawiska, w którym zbiór danych użyty do zbudowania i trenowania modelu uczenia maszynowego lub sztucznej inteligencji nie odzwierciedla w odpowiedni sposób całej populacji, której ma służyć.

Błąd historyczny: Błędy te pochodzą z błędów historycznych w danych treningowych i są odtwarzane w systemach AI. Na przykład, jeśli w przeszłości praktyki rekrutacyjne były stronnicze ze względu na grupę demograficzną, to narzędzie AI odtworzyłoby te stronnicze zachowania, rekomendując kandydatów, którzy pod względem cech demograficznych przypominali te, które mają dane miejsce.

Błąd etykiety: Kiedy etykiety przypisane konkretnym instancjom danych okazują się stronnicze lub niepoprawne. Na przykład, zbiór danych utworzony przez pojedynczą grupę demograficzną: dane dotyczące zdrowia w przypadku analizy sentymentu. Model uczyłby się niesprawiedliwych skojarzeń dla niektórych sentymentów.

Błąd agregacji: Dane zebrane lub zagregowane w sposób, który pomija istotne różnice między podgrupami. Na przykład, zagregowane dane dotyczące zdrowia mogą maskować regionalne lub demograficzne różnice w wynikach zdrowotnych.

Błąd temporalny: Występuje, gdy rodzaj danych zmienia się w czasie, ale model sztucznej inteligencji nie dostosowuje swojego działania do tych zmian. Na przykład, systemy rekomendacji oparte na historycznych danych historycznych bez uwzględniania zmieniających się preferencji użytkowników mogą ostatecznie rekomendować te same produkty w każdym przypadku.

Błąd amplifikacji: Termin ten odnosi się do sytuacji, w których pewne punkty danych lub wzorce otrzymują więcej uwagi ze strony danych treningowych modelu sztucznej inteligencji, niż pozwalałoby na to w pełni rozwinięte podejście. Na przykład, algorytm rankingowy wyszukiwarki zwiększa popularność popularnych stron internetowych, dodatkowo je promując poprzez pętle sprzężenia zwrotnego. Błąd społeczno-ekonomiczny: Można go przypisać sposobowi, w jaki system sztucznej inteligencji (AI) odtwarza nierówności społeczno-ekonomiczne występujące w społeczeństwie. Przykładem może być model scoringu kredytowego, który stawia osoby o niskich dochodach w niekorzystnej sytuacji, ponieważ został wytrenowany z wykorzystaniem danych korelujących wiarygodność kredytową ze statusem społeczno-ekonomicznym. Błąd kulturowy: Występuje, gdy dane treningowe są obciążone uprzedzeniami wynikającymi z perspektywy kulturowej lub uprzedzeń osoby zbierającej dane lub adnotującej. Jeśli model AI jest trenowany na danych pochodzących tylko z jednej kultury lub języka, może mieć trudności ze zrozumieniem lub wygenerowaniem tekstu w innych kontekstach kulturowych.

Aby przeciwdziałać tym błędom, programiści muszą projektować swoje procesy i modele sztucznej inteligencji w sposób wolny od błędów, począwszy od gromadzenia danych, przetwarzania, przez trenowanie modeli, aż po testowanie. Takie podejście może obejmować takie elementy, jak metody budowania zbiorów danych, dbanie o uczciwość algorytmów oraz opracowywanie metryk do oceny uczciwości.

Błąd w AI/ML

Termin błąd w AI/ML odnosi się do systematycznych błędów lub niesprawiedliwości występujących w rozwoju, wdrażaniu lub stosowaniu algorytmów AI i ML. Błąd może pojawić się na niemal wszystkich etapach cyklu życia AI/ML, od gromadzenia danych do wdrożenia. W systemach AI/ML może występować kilka rodzajów błędów.

Błąd próbkowania: Pojawia się, gdy dane szkoleniowe nie charakteryzują dokładnie populacji, którą mają modelować. Prowadzi to do błędnych szacunków lub klasyfikacji.

Błąd algorytmiczny: Wynika ze strategii i wdrażania samych algorytmów uczących się. Niektóre algorytmy mogą być z natury stronnicze w kierunku określonych wyników, zwłaszcza jeśli dane są przetwarzane w określony sposób.

Błąd pomiaru: Jest to błąd w procesie pomiaru. Jeśli pewne grupy byłyby systematycznie niedoreprezentowane lub błędnie reprezentowane w danych, prowadziłoby to do błędnych wyników.

Błąd historyczny: Odzwierciedla błąd występujący w danych historycznych, na których trenowany jest model. W przypadku decyzji z przeszłości obciążonych błędami historycznymi, model prawdopodobnie będzie je propagował, chyba że zostaną podjęte działania naprawcze.

Błąd reprezentacji: Występuje w przypadku braku reprezentacji jednostek, grup lub określonej klasy obecnej w zbiorze danych, co powoduje zniekształcenie ich systemów.

Błąd oceny: Występuje w postaci zestawu kryteriów pomiaru skuteczności systemu AI/ML, które same w sobie są obarczone błędami. Skutkuje to wadliwą oceną wydajności systemu.

Aby przeciwdziałać stronniczości w AI/ML, konieczne jest holistyczne podejście na wszystkich etapach. Różnorodność i jakość danych szkoleniowych, wraz z technikami takimi jak algorytmy uwzględniające uczciwość i audyty stronniczości, muszą być częścią arsenału. Przejrzystość i możliwość wyjaśnienia umożliwiają wykrywanie i rozwiązywanie stronniczości, a zróżnicowany zespół i zaangażowanie interesariuszy zapewniają całościowy obraz rozwoju systemu. Ramy regulacyjne i wytyczne etyczne definiują standardy uczciwości i niedyskryminacji. Monitorowanie w warunkach rzeczywistych, wraz z ciągłym doskonaleniem opartym na informacjach zwrotnych, pomaga minimalizować stronniczość, zapewniając, że systemy AI/ML są wdrażane i wykorzystywane w sposób etyczny i sprawiedliwy w różnych społeczeństwach, budując zaufanie i maksymalizując korzyści społeczne.

Błąd w ocenie ludzkiej

Źródła błędu w ocenie ludzkiej w generatywnej sztucznej inteligencji mogą być różnorodne i mogą wystąpić na dowolnym etapie procesu rozwoju sztucznej inteligencji . Oto kilka ważnych aspektów, które należy wziąć pod uwagę:

Błąd danych szkoleniowych: Jeśli dane szkoleniowe dla generatywnych modeli sztucznej inteligencji nie są wystarczająco heterogeniczne lub reprezentatywne, wyniki będą obarczone błędem. Załóżmy na przykład, że model językowy jest trenowany na tekście bardziej reprezentatywnym dla określonego segmentu populacji lub regionu. W związku z tym model miałby trudności z opracowaniem dokładnych i rzetelnych odpowiedzi dla innych.

Adnotacje dotyczące błędów: Adnotacje są zazwyczaj tworzone przez osoby uczestniczące w procesie szkolenia. Ludzie mogą czasami personalizować lub dyskryminować w etykietowaniu działań na podstawie osobistych preferencji, kultury lub innych uprzedzeń społecznych. Zmienia to sposób, w jaki model będzie interpretował generowaną treść.

Błąd w ocenie: Generatywna sztuczna inteligencja może być narażona na błąd ze względu na sposób, w jaki jest oceniana przez recenzentów. Na przykład, jeśli oceny wyników opierają się na osobistych preferencjach, pochodzeniu kulturowym lub stereotypach społecznych, może to prowadzić do niesprawiedliwych ocen.

Błąd precyzyjnego dostrajania: Zbiory danych i parametry precyzyjnego dostrajania używane w generatywnym modelu sztucznej inteligencji mogą powodować jego stronniczość. Na przykład, wyniki takiego precyzyjnie dostrojonego modelu będą obarczone błędem, jeśli zbiór danych reprezentuje głównie jedną konkretną grupę demograficzną.

Błąd związany z udziałem człowieka w procesie generowania: W przypadkach takich jak moderacja treści i platformy generowania tekstu, ludzie uczestniczący w procesie generowania mogą nieumyślnie, świadomie lub nieświadomie, wzmacniać lub wyolbrzymiać błędy modelu. W związku z tym ich decyzje o akceptacji/odrzuceniu wyników w większym stopniu odzwierciedlają ich błędy niż wiarygodność samego wyniku.

Łagodzenie błędów w ocenie ludzkiej obejmuje zarówno interwencje techniczne, takie jak metody wykrywania i łagodzenia błędów w trakcie rozwoju modelu, jak i szersze działania mające na celu zwiększenie różnorodności i inkluzywności badań i rozwoju AI. Twórcy oprogramowania muszą dostrzegać te błędy i rozważać podejmowanie działań w celu łagodzenia ich skutków na każdym etapie cyklu życia rozwoju AI.

W tej sekcji, za pomocą Tabeli, zilustrowano, że różne formy uprzedzeń pojawiły się już w sztucznej inteligencji, często w formie założeń i innych pozytywnych dowodów na wpływ tej predyspozycji na systemy sztucznej inteligencji, podkreślając w ten sposób potrzebę właściwej oceny i przeciwdziałania. Tabela zawiera przykłady różnych form uprzedzeń i inne pozytywne dowody na wpływ takich uprzedzeń na systemy sztucznej inteligencji.

Kategoria stronniczości: Opis: Scenariusze

Strategia danych: Tendencja do odzwierciedlania i utrwalania uprzedzeń społecznych obecnych w treściach wejściowych.: Model języka stronniczego, trenowany na stronniczych tekstach internetowych, może generować wulgarny lub rasistowski język.

Strategia sztucznej inteligencji/uczenia maszynowego: Istniejąca wcześniej tendencja do powielania i wzmacniania istniejących uprzedzeń społecznych w generowanych danych wyjściowych.: System rozpoznawania twarzy trenowany na danych jednej grupy etnicznej mógłby błędnie klasyfikować twarze osób innej grupy etnicznej częściej.

Strategia recenzji ludzkiej: Subiektywizm ludzkich ewaluatorów może również ujawniać osobiste uprzedzenia podczas oceny generowanych treści.: Treści generowane przez takie filtry mogą z natury rzeczy oznaczać, że nie staną się one w pełni bezstronną funkcją lub będą oceniane inaczej przez recenzentów, w zależności od ich przekonań kulturowych lub osobistych, wpływając w ten sposób na ich postrzeganą jakość.

Wpływ uprzedzeń

Postęp w dziedzinie sztucznej inteligencji (AI) niesie ze sobą zarówno korzyści, jak i zagrożenia. W szczególności uprzedzenia pogłębiają nierówności, utrwalają stereotypy i sprzyjają dyskryminacji osób spoza centralnych struktur władzy. Rodzi to obawy etyczne dotyczące niesprawiedliwości i uprzedzeń wynikających z ryzyka związanego z uprzedzeniami w AI, co może prowadzić do negatywnych skutków. Zajęcie się tymi kwestiami w tym kontekście wiąże się z obawami etycznymi związanymi z niesprawiedliwością i uprzedzeniami. W erze postcyfrowej rewolucja, łagodzenie problemów związanych z uprzedzeniami w AI budzi takie obawy .

Negatywne skutki

Uprzedzenia w sztucznej inteligencji (AI) mają ogromny i negatywny wpływ na jednostki i społeczeństwo poprzez reprodukcję dyskryminacji i pogłębianie status quo w następujący sposób:

* Algorytmy dyskryminacji są tworzone w oparciu o dane obarczone stronniczością. Na przykład, jeśli istnieją dane historyczne dotyczące algorytmów rekrutacyjnych, które są stronnicze wobec pewnych grup demograficznych, takich jak płeć lub rasa, to uprzedzenia te zostaną spotęgowane przez algorytm, który będzie dyskryminował kandydatów z grup uprzywilejowanych i poświęcał niezauważoną uwagę osobom z grup upośledzonych.

* Wspieranie stereotypów: Jest więcej niż prawdopodobne, że stronnicze systemy AI będą wzmacniać istniejące stereotypy dotyczące niektórych osób i utrzymywać je w marginalizacji i ucisku. Ponadto stronnicze algorytmy rozpoznawania twarzy zakładałyby, że osoby o innym kolorze skóry lub pochodzeniu rasowym lub etnicznym są bardziej narażone, przewidując w ten sposób coś, co mogłoby ostatecznie wzmocnić negatywne stereotypy i prowadzić do większej kontroli lub niesprawiedliwego traktowania.

* Odrzucenie i uprzedzenia: Stronnicze systemy sztucznej inteligencji pomijają wszystkich poza uciskanymi. Na przykład, stosowanie niskiej oceny zdolności kredytowej może czasami oznaczać osoby należące do gospodarstw domowych o niskich dochodach lub mniejszości jako przegranych.

* Podważanie równego dostępu: Głównie poprzez stronnicze wprowadzanie sztucznej inteligencji (AI) możliwe jest niesprawiedliwe traktowanie takich możliwości jak edukacja, zatrudnienie i opieka zdrowotna. Na przykład, ponieważ stronniczy algorytm predykcyjny policji ma tendencję do silniejszego angażowania się w społeczności mniejszościowe, skutkuje to wzmożoną inwigilacją, nękaniem i bezpodstawnymi aresztowaniami – wszystko w imię walki z przestępczością.

* Utrata zaufania: Od momentu, gdy systemy AI generują stronnicze wyniki, zaufanie jest tracone, szczególnie gdy są wykorzystywane w kluczowych miejscach. Utrata zaufania może być zbyt daleko idąca, ponieważ powoduje zmniejszoną współpracę z organami ścigania, mniejszy udział społeczeństwa i wzrost napięć społecznych.

* Stronniczość w sztucznej inteligencji utrwala nierówności istniejące w sektorze technologicznym. Algorytmy stronnicze mogą być opracowywane przez zespoły składające się z jednorodnych, pozbawionych różnorodności grup ludzi, które z większym prawdopodobieństwem utrwalają i utrwalają uprzedzenia w swoich zespołach, co prowadzi do braku różnorodności i aneksji w rozwoju technologii.

* Uprzedzenia w sztucznej inteligencji są kwestią prawną i etyczną, a w szczególności w kontekście dyskryminacji i sprawiedliwości. Organizacje mogą stawić czoła pozwom sądowym za dyskryminację lub utracić ważnego klienta, jeśli wykorzystywane przez nie systemy sztucznej inteligencji okażą się stronnicze. Może to prowadzić do pozwów sądowych, a także kontroli ze strony organów regulacyjnych.

Zatem system sztucznej inteligencji może wykazywać liczne zniekształcenia, które wymagają wysiłków ze strony decydentów, technologów i całego społeczeństwa. Konieczne jest wdrożenie solidnych testów i walidacji w celu wykrywania i łagodzenia stronniczości przed jej wdrożeniem. Ponadto, wdrażane systemy sztucznej inteligencji powinny charakteryzować się przejrzystością i rozliczalnością, a także ciągłymi działaniami edukacyjnymi i podnoszącymi świadomość, które zapewnią, że społeczeństwo będzie informowane i będzie miało odpowiednie uprawnienia, co pozwoli na zapewnienie sprawiedliwych i równych systemów sztucznej inteligencji.

Wnioski etyczne stronniczej sztucznej inteligencji (AI)

Konsekwencje stronniczej sztucznej inteligencji (AI) oznaczają zatem utrwalanie dyskryminacji grup marginalizowanych, co jest bardzo istotne, szczególnie w kontekście utrwalania nierówności w dostępie do opieki zdrowotnej i zatrudnienia. Pogłębi to istniejące nierówności. Twórcy oprogramowania, firmy i rządy ponoszą odpowiedzialność za zagwarantowanie sprawiedliwości i przejrzystości w projektowaniu i wdrażaniu AI. Brak reakcji na stronnicze rezultaty budzi nie tylko podejrzliwość opinii publicznej wobec technologii, ale także grozi uniemożliwieniem realizacji korzyści płynących z AI. Wymaga to większej odpowiedzialności wszystkich interesariuszy, którzy będą

zaangażowani w rozwiązywanie kwestii etycznych. Ponadto konieczne jest opracowanie jasnych wytycznych i ram etycznych, aby korzystanie z AI i jego rozwój były sprawiedliwe, przejrzyste i rozliczalne. Należy również podjąć dalsze kroki w celu stworzenia dyskusji kluczowych dla wdrażania AI i zbudowania społeczeństwa na rzecz odpowiedzialnego i etycznego korzystania z AI. Uprzedzenia w sztucznej inteligencji (AI) wiążą się z poważnymi problemami etycznymi, od moralnych po wielorakie:

* Uprzedzenia w sztucznej inteligencji utrwalają niesprawiedliwe traktowanie, a niektóre z ich skutków dotykają w szczególności członków grup marginalizowanych.

* Nieprzejrzyste systemy AI ograniczają rozliczalność i przejrzystość.

* Uprzedzenia w sztucznej inteligencji mogą potencjalnie naruszać prywatność poprzez ujawnianie poufnych informacji.

* Uprzedzenia w sztucznej inteligencji (AI) mogą prowadzić do zmniejszenia zaufania do technologii i instytucji.

* Uprzedzenia w sztucznej inteligencji mogą nie być przestrzegane z powodu uprzedzeń w sztucznej inteligencji.

* Uprzedzenia w sztucznej inteligencji muszą być proaktywnie łagodzone, aby im zapobiegać.

* Edukacja i świadomość są niezbędne do wspierania odpowiedzialnych praktyk w zakresie sztucznej inteligencji.

Metodologie łagodzenia uprzedzeń w AI

Łagodzenie uprzedzeń w systemach AI jest ważne dla sprawiedliwości i równości w wielu kontekstach. Pomimo tych problemów, interdyscyplinarne badania, których celem będzie opracowanie przejrzystych praktyk w rozwoju i wdrażaniu AI, są niezbędne do pokonania tych przeszkód. Rozważania etyczne są istotne w przezwyciężaniu problemów związanych z uprzedzeniami w AI, biorąc pod uwagę trudne kwestie, z którymi należy się zmierzyć, dotyczące priorytetu typów uprzedzeń i grup dotkniętych. Pomimo istniejących wyzwań, opracowanie holistycznych strategii łagodzenia poprzez współpracę informatyków, etyków i socjologów ma kluczowe znaczenie dla zapewnienia, że systemy AI przyczyniają się do dobrobytu wszystkich podmiotów i społeczeństwa. Postęp społeczny i poprawa dobrostanu człowieka muszą być priorytetem, aby technologie budowały zaufanie do siebie. Może to obejmować wstępne przetwarzanie danych, być może z wykorzystaniem technik takich jak augmentacja danych, selekcja cech i dystrybucja klas przed ich wykorzystaniem w modelach AI. Jednak jednym z wyzwań, jakie stwarza wstępne przetwarzanie danych, jest dostępność zróżnicowanych i reprezentatywnych danych szkoleniowych. W tym celu błędy stwierdzone w danych szkoleniowych mogą rozprzestrzeniać się na modele sztucznej inteligencji (AI) i prowadzić do wyników obarczonych błędami, takimi jak niesprawiedliwość lub dyskryminacja. Błędy te są również trudne do wykrycia i ograniczenia w danych, ponieważ mogą być ukryte w kontekście. Wybór modelu obejmuje proces wyboru najlepszego modelu lub architektury do trenowania modeli AI. Różne modele różnią się pod względem zdolności do utrwalania błędów; dlatego wybór najlepszego modelu pomaga ograniczyć błędy w systemach AI. Jednak nawet najnowocześniejsze modele mogą być obarczone błędami i zawsze będą istniały kompromisy między uczciwością a dokładnością. Modele zoptymalizowane pod kątem dokładności mogą, nieoczekiwanie, utrwalać błędy w danych szkoleniowych. Dlatego wybór modelu powinien być dokonywany rozważnie, w zależności od domeny docelowej i możliwych błędów u niektórych zaangażowanych interesariuszy.

Postprocessing obejmuje decyzje dotyczące technik stosowanych po tym, jak model dokona prognoz w celu zmniejszenia uprzedzeń i skorygowania wyników. Przykładami takiego postprocessingu są korekta uprzedzeń, postprocessing uwzględniający rzetelność oraz modele interpretowalne. Czasami, chociaż techniki postprocessingu umożliwiają bezpośrednie zastosowanie do wyników modelu, nie zawsze pomagają w redukcji uprzedzeń. Ponieważ zawsze istnieje ryzyko wprowadzenia nowych uprzedzeń lub niezamierzonych konsekwencji poprzez interwencje postprocessingowe, istnieje również ryzyko ich wystąpienia. Względy etyczne są również jednym z najważniejszych punktów w walce z uprzedzeniami w sztucznej inteligencji. Priorytetowe traktowanie różnych kategorii uprzedzeń i różnych grup w działaniach moderacyjnych wiąże się ze złożonymi problemami etycznymi. Na przykład, czy główny nacisk należy położyć na redukcję uprzedzeń wobec grup historycznie marginalizowanych, czy na próbę redukcji uprzedzeń, które mogą faworyzować określone grupy uprzywilejowane? Co więcej, istnieje potrzeba konkurencji między konkurującymi zasadami etycznymi: rzetelnością, przejrzystością, prywatnością i użytecznością. Uprzedzenia w sztucznej inteligencji pogłębiają istniejące nierówności i utrwalają niesprawiedliwą dyskryminację, co ma niekorzystne konsekwencje dla grup marginalizowanych. Ponadto, stronnicza sztuczna inteligencja może podważyć zaufanie do systemów sztucznej inteligencji oraz obniżyć ich legalność i akceptowalność. Dalsze badania i rozwój metod łagodzenia skutków są zatem niezwykle ważne, aby stawić czoła tym wyzwaniom w taki sposób, aby systemy sztucznej inteligencji mogły być wykorzystywane z korzyścią dla wszystkich. Co więcej, konieczne są przejrzyste i odpowiedzialne praktyki w zakresie rozwoju i wdrażania sztucznej inteligencji, które budują zaufanie do systemów sztucznej inteligencji. Wreszcie, zwalczając uprzedzenia w sztucznej inteligencji, możemy osiągnąć całkowite korzyści płynące z technologii sztucznej inteligencji na rzecz postępu społecznego, równości i poprawy dobrobytu ludzi. Tabela 2 przedstawia wyzwania związane z mechanizmami łagodzenia uprzedzeń w sztucznej inteligencji wraz z odpowiednimi rozwiązaniami.

Podejścia: Ograniczenia: Rozwiązania

Wstępne przetwarzanie danych:

- Ograniczona różnorodność

- Subtelne uprzedzenia

- Nierównowaga danych:

- Augmentacja danych

- Narzędzia do wykrywania uprzedzeń

- Techniki łagodzenia nierównowagi klas

Wybór modelu:

- Kompromis w zakresie dokładności rzetelności

- Uprzedzenia algorytmiczne

- Specyficzność domeny:

- Metryki rzetelności

- Techniki łagodzenia uprzedzeń

- Współpraca i wiedza specjalistyczna

Decyzje po przetworzeniu:

- Skuteczność

- Ryzyko wprowadzenia uprzedzeń

- Interpretowalność:

- Ramy ewaluacji

- Przejrzystość i audytowalność

- Wyjaśnialna AI

Sprawiedliwość: wskaźniki i strategie łagodzenia

Sprawiedliwość w AI, czyli obecność stronniczości lub dyskryminacji w systemach AI, to kolejny gorąco dyskutowany temat zarówno w literaturze akademickiej, jak i biznesowej. Kwestia ta jest ściśle związana z różnymi rodzajami uprzedzeń, które mogą się pojawić. Proponowane typy obejmują sprawiedliwość grupową, indywidualną i kontrfaktyczną. Chociaż sprawiedliwość i stronniczość są ze sobą ściśle powiązane, sprawiedliwość jest zamierzonym celem, podczas gdy stronniczość może być niezamierzonym rezultatem. Osiągnięcie sprawiedliwości w AI wymaga przemyślanego rozważenia kontekstu i interesariuszy. Jako przykłady korzyści płynące z sprawiedliwości w AI zilustrowano w praktyce . Kluczowe zagadnienia dotyczące sprawiedliwości w AI obejmują:

* Wykrywanie i łagodzenie uprzedzeń

* Przejrzystość i możliwość wyjaśnienia

* Zróżnicowaną reprezentację

* Odpowiedzialność i zarządzanie

* Równość w dostępie i wpływie

Źródła sprawiedliwości

Poniżej opisano różne rodzaje sprawiedliwości :

Sprawiedliwość grupowa: Sprawiedliwość decyzji we wszystkich grupach demograficznych.

Sprawiedliwość indywidualna: Osoby o podobnych cechach powinny być traktowane w podobny sposób.

Sprawiedliwość przyczynowo-skutkowa: Należy rozważyć, jak decyzja wpłynie na różne grupy demograficzne.

Sprawiedliwość czasowa: Należy przestrzegać tych samych zasad, aby sprawiedliwość mogła być utrzymana przez określony czas.

Sprawiedliwość algorytmiczna: Ograniczenia sprawiedliwości muszą być uwzględnione w algorytmach.

Sprawiedliwość procesowa: Przejrzystość i rozliczalność w procesach decyzyjnych.

Sprawiedliwość domenowa: Kwestie sprawiedliwości w kontekście konkretnych domen.

Tego rodzaju uczciwość często wiąże się z kompromisami i koniecznością uwzględnienia zasad etycznych, wymogów prawnych i wartości społecznych, co wymaga stałego uświadomienia sobie, że uczciwość jest w istocie złożoną i wielowymiarową koncepcją, która wymaga intensywnych badań, debat i partnerstwa, aby mogła zostać wdrożona w praktyce.

Mierniki sprawiedliwości w AI

Sprawiedliwość w AI to bardzo skomplikowana i trudna do monitorowania koncepcja, a nawet gdybyśmy uzgodnili mierniki sprawiedliwości, eksperci w dziedzinie etyki praktyk nie byliby w stanie stworzyć żadnych ogólnie akceptowalnych mierników. W praktyce do oceny sprawiedliwości w systemach AI stosuje się wiele miar i podejść sprawiedliwości:

Parytet demograficzny: Odnosi się to do stopnia jednolitości prognoz lub wyników w różnych grupach demograficznych reprezentowanych przez takie atrybuty, jak rasa, płeć i wiek. Gdyby osiągnięto parytet demograficzny prognoz lub decyzji, oznaczałoby to, że system AI formułuje prognozy lub decyzje bez względu na jakiekolwiek atrybuty demograficzne.

Równość szans: Polega ona na badaniu konkretnych wskaźników prawdziwie pozytywnych (poprawnych prognoz) dla różnych grup demograficznych, tak aby prawdopodobieństwo prawidłowej identyfikacji lub klasyfikacji było równe.

Wyrównane szanse: Polega na bezstronności modelu. Zapewnia to równowagę wskaźników błędów modelu we wszystkich grupach.

Parzystość predykcyjna: Parzystość predykcyjna mierzy rozkład szacunków sporządzonych przez prototyp w różnych klastrach w celu zapewnienia sprawiedliwości. Sprawdza, czy proporcja pozytywnych prognoz (np. zatwierdzenia wniosku o pożyczkę) jest podobna w różnych grupach demograficznych.

Równość traktowania: Sprawdza, czy osoby z różnych grup demograficznych otrzymają takie samo traktowanie lub wynik od systemu AI, niezależnie od ich cech demograficznych.

Sprawiedliwość kontrfaktyczna: Sprawdza, czy osoby otrzymałyby taki sam wynik, gdyby ich cechy demograficzne były inne, przy zachowaniu niezmienności pozostałych cech.

Sprawiedliwość przyczynowa: Sprawdza sprawiedliwość relacji między cechami wejściowymi, decyzjami podjętymi przez system AI i wynikami w procesie decyzyjnym.

Sprawiedliwość intersekcjonalna: Analizuje, w jaki sposób wszystkie brane pod uwagę dane wejściowe prowadzą do wyniku, biorąc pod uwagę przecięcia cech demograficznych, aby upewnić się, że osoby o zbieżnych tożsamościach nie będą nieproporcjonalnie narażone na stronniczość w systemie AI.

Wskaźniki łagodzenia uprzedzeń: To nie bezpośrednia miara, ale skuteczność technik redukcji uprzedzeń stosowanych w systemach sztucznej inteligencji umożliwi pośredni pomiar sprawiedliwości. Ograniczenia sprawiedliwości, trening uwzględniający uprzedzenia oraz korekty algorytmiczne to podejścia stosowane w celu redukcji uprzedzeń i zwiększenia sprawiedliwości w systemach sztucznej inteligencji.

Te wskaźniki i podejścia dotyczące sprawiedliwości stanowią ramy dla oceny sprawiedliwości w systemach AI, jednak muszą być modyfikowane, aby pasowały do konkretnych kontekstów i dziedzin.

Ponadto, ocena sprawiedliwości często wymaga połączenia analizy ilościowej, oceny jakościowej i opinii interesariuszy, aby właściwie zrozumieć i rozwiązać problemy związane ze sprawiedliwością w systemach AI.

Metodologie ograniczania błędów w celu zapewnienia uczciwości AI

Stosuje się kilka podejść do zapewnienia uczciwości w ograniczaniu błędów i równych wyników w systemach AI . Oto niektóre z nich:

Wykrywanie i pomiar błędów: Zanim jeszcze zostaną one wyeliminowane, niezbędne jest ich wykrycie i pomiar. Obejmuje to ocenę zbiorów danych, algorytmów i wyników pod kątem błędów wynikających z rasy, płci, wieku lub innych wrażliwych cech.

Augmentacja danych: Obejmuje to wstępne przetwarzanie danych w celu ograniczenia błędów i promowania zróżnicowanego uwzględnienia w zbiorze danych.

Techniki obejmują:

- Równoważenie niedostatecznie reprezentowanych grup poprzez generowanie danych syntetycznych.

- Stosowanie uczciwych technik próbkowania, które zapewniają uwzględnienie wszystkich grup demograficznych w danych.

- Selekcję i modyfikację cech.

- Uczenie się uczciwej reprezentacji: uczenie się reprezentacji danych, które redukują błędy.

- Szkolenie w zakresie modelu sprawiedliwości: włączanie ograniczeń lub kar dotyczących sprawiedliwości w procesie szkolenia modeli w celu uniknięcia dyskryminacji.

- Szkolenie w zakresie podejścia adwersarskiego: szkolenie modeli w zakresie przeciwdziałania atakom adwersarskim w celu ujawnienia i ograniczenia uprzedzeń.

Interpretowalna i wyjaśnialna sztuczna inteligencja: Najlepsze modele muszą oferować kompletne i łatwe do zrozumienia wyjaśnienia, w jaki sposób podjęły swoje decyzje, zwiększając w ten sposób przejrzystość i rozliczalność, pomagając znacznie skuteczniej identyfikować i korygować uprzedzenia.

Ciągły monitoring i audyt: Obejmuje to:

Metodologie ograniczania błędów w celu zapewnienia uczciwości AI

Stosuje się kilka podejść do zapewnienia uczciwości w ograniczaniu błędów i równych wyników w systemach AI . Oto niektóre z nich:

Wykrywanie i pomiar błędów: Zanim jeszcze zostaną one wyeliminowane, niezbędne jest ich wykrycie i pomiar. Obejmuje to ocenę zbiorów danych, algorytmów i wyników pod kątem błędów wynikających z rasy, płci, wieku lub innych wrażliwych cech.

Augmentacja danych: Obejmuje to wstępne przetwarzanie danych w celu ograniczenia błędów i promowania zróżnicowanego uwzględnienia w zbiorze danych.

Techniki obejmują:

- Równoważenie niedostatecznie reprezentowanych grup poprzez generowanie danych syntetycznych.

- Stosowanie uczciwych technik próbkowania, które zapewniają uwzględnienie wszystkich grup demograficznych w danych.

- Selekcję i modyfikację cech.

- Uczenie się uczciwej reprezentacji: uczenie się reprezentacji danych, które redukują błędy.

- Szkolenie w zakresie sprawiedliwego modelu: uwzględnianie ograniczeń lub kar dotyczących uczciwości w procesie szkolenia modeli w celu uniknięcia dyskryminacji.

- Szkolenie w zakresie obrony: szkolenie modeli przed atakami ze strony przeciwnika w celu ujawnienia i ograniczenia uprzedzeń.

Interpretowalna i wyjaśnialna sztuczna inteligencja: Najlepsze modele muszą oferować kompletne i łatwe do zrozumienia wyjaśnienia, w jaki sposób podjęły swoje decyzje, zwiększając w ten sposób przejrzystość i rozliczalność, pomagając w znacznie skuteczniejszym identyfikowaniu i korygowaniu uprzedzeń.

Ciągły monitoring i audyt: Obejmuje on:

- Monitorowanie wydajności modelu: stałe monitorowanie wydajności modelu w różnych grupach demograficznych.

- Audyty uprzedzeń: okresowy przegląd zbiorów danych i modeli pod kątem uprzedzeń i podejmowanie działań korygujących w razie potrzeby.

- Interwencje regulacyjne i polityczne: wdrażanie zasad i polityk w celu zapewnienia, że systemy sztucznej inteligencji działają w sposób sprawiedliwy i odpowiedzialny. Może to obejmować ramy prawne, standardy branżowe i wytyczne dotyczące etycznego rozwoju i wdrażania sztucznej inteligencji.

* Przejrzystość i rozliczalność algorytmów: Mechanizmy powinny umożliwiać wyjaśnienie ich działania, aby umożliwić jasne zrozumienie określania i ustalania konsekwencji poprzez zapewnienie przejrzystości w podejmowaniu decyzji i osiąganych rezultatów .

* Rozważania etyczne: Przy opracowywaniu systemów AI należy uwzględnić szersze implikacje etyczne dotyczące ważnej roli, jaką systemy AI odgrywają w społecznościach narażonych na ataki.

* Edukacja i świadomość: Społeczeństwo oraz interesariusze ze świata biznesu i nauki muszą zrozumieć znaczenie uczciwości w AI oraz praktyk zapewniających etyczny rozwój AI .

Połączenie tych metodologii może pomóc ograniczyć uprzedzenia i zapewnić bardziej sprawiedliwe wyniki dla różnych grup społecznych.

Wnioski

W niniejszym rozdziale omówiono ewolucję uprzedzeń w systemach sztucznej inteligencji (AI) i uczenia maszynowego (ML) oraz ich ogromny wpływ na społeczeństwo; szczegółowo wyjaśniono narastające problemy związane z uprzedzeniami generatywnymi w AI. Podkreślono, że te zaawansowane narzędzia obliczeniowe, które wkrótce zostaną udostępnione, mogą wzmocnić istniejące uprzedzenia, przede wszystkim dotyczące płci i rasizmu, ale także innych czynników kulturowych. W szczególności w rozdziale przeanalizowano kilka przypadków stronniczych systemów AI, koncentrując się głównie na złożoności generatywnej AI. Pokazano, jak ważne jest holistyczne podejście do wykrywania i redukcji uprzedzeń w całym procesie rozwoju AI. Istnieje potrzeba, aby przyszłe badania nad sprawiedliwością i uprzedzeniami w AI i uczeniu maszynowym (ML) koncentrowały się bardziej na danych szkoleniowych i subtelnych kwestiach uprzedzeń nieodłącznie związanych z modelami generatywnymi, zwłaszcza tymi generującymi dane i treści syntetyczne. Istnieje pilna potrzeba opracowania polityk i obowiązków,

które umożliwiłyby odpowiedzialne wdrażanie AI i uczenia maszynowego. Powinny one obejmować otwarte dane szkoleniowe, wybór modeli i procesy generowania. Podobnie jak w przypadku uprzedzeń, należy powołać zróżnicowane zespoły pracujące nad rozwojem i oceną sztucznej inteligencji, aby zapewnić większą różnorodność i punkty widzenia, które pomogą w skutecznym wykrywaniu i rozwiązywaniu uprzedzeń.