Matematyka uprzedzeń: granice sprawiedliwości algorytmicznej w świetle książki Raising AI Chiraga Shaha

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje rozdźwięk między techniczną poprawnością modeli AI a sprawiedliwością ich rezultatów społecznych. Autor wskazuje, że wysoka precyzja statystyczna nie eliminuje uprzedzeń, gdyż modele często opierają się na wadliwych danych historycznych lub błędnej operacjonalizacji pojęć (brak trafności konstruktu). Tekst wyjaśnia, że systemy AI mogą utrwalać nierówności poprzez specyficzny dobór zmiennych i funkcji straty, co prowadzi do tzw. matematyki uprzedzeń. Omówiono również typologię stronniczości algorytmicznej – od historyczną i reprezentacyjnej po pomiarową i wdrożeniową – podkreślając, że neutralność techniczna nie jest tożsama z neutralnością poznawczą czy etyczną.

The article analyzes the gap between the technical correctness of AI models and the fairness of their social outcomes. The author points out that high statistical precision does not eliminate bias, as models often rely on flawed historical data or incorrect operationalization of concepts (lack of construct validity). The text explains that AI systems can perpetuate inequalities through specific selections of variables and loss functions, leading to the so-called mathematics of bias. It also discusses a typology of algorithmic bias—from historical and representational to measurement and deployment bias—emphasizing that technical neutrality is not synonymous with cognitive or ethical neutrality.

Artykuł stanowi krytyczną analizę złudzenia, jakoby techniczna precyzja algorytmów sztucznej inteligencji była tożsama z obiektywizmem i sprawiedliwością społeczną. Autor dowodzi, że systemy AI mogą być matematycznie bezbłędne, a jednocześnie systematycznie dyskryminować określone grupy osób poprzez tzw. „matematykę

uprzedzeń". Główna teza tekstu zakłada, że sprawiedliwość algorytmiczna nie jest problemem programistycznym, lecz złożonym splotem kwestii statystycznych, instytucjonalnych i aksjologicznych. Analizując zjawiska takie jak paradoks Simpsona, błędy selekcji czy stronniczość zmiennych zastępczych (proxy bias), tekst wykazuje, że AI często nie tyle rozwiązuje problemy, co utrwala historyczne nierówności pod maską neutralności danych. Znaczenie tej refleksji polega na wezwaniu do przejścia od naiwnego technokratyzmu ku „wychowaniu" sztucznej inteligencji – procesowi, w którym nad matematyczną optymalizacją dominuje świadome definiowanie celów społecznych oraz rygorystyczna kontrola nad etycznymi konsekwencjami automatyzacji decyzji o ludzkim życiu.

SŁOWA KLUCZOWE

sprawiedliwość algorytmiczna | matematyka uprzedzeń | stronniczość reprezentacyjna | paradoks Simpsona | collider bias | wyniki kontrfaktyczne | parytet demograficzny | class imbalance | trafność konstruktu | epistemologia danych | funkcja straty | wnioskowanie przyczynowe | niesprawiedliwość hermeneutyczna | metryka F1 | podstawa Bayesa

Techniczna poprawność modelu nie gwarantuje sprawiedliwości społecznej

Matematyka uprzedzeń i cicha produkcja nierówności. Epistemologia danych, statystyczne paradoksy i granice algorytmicznej sprawiedliwości

Jednym z najtrudniejszych wyzwań współczesnej sztucznej inteligencji jest rozdźwięk między matematyczną poprawnością procedury obliczeniowej a społeczną sprawiedliwością jej rezultatów. Zjawisko to podważa powszechne przekonanie, że zastosowanie zaawansowanej statystyki automatycznie eliminuje ludzkie uprzedzenia, arbitralność czy subiektywizm. Algorytm może bowiem bezbłędnie wykonywać wszystkie operacje i osiągać wysokie wskaźniki jakości, a jednocześnie systematycznie pogarszać sytuację określonych grup osób. Nie musi to wynikać z błędu programistycznego ani intencjonalnej dyskryminacji wpisanej w kod.

Źródłem problemu bywa sposób definiowania celu, dobór populacji treningowej, konstrukcja zmiennej objaśnianej, struktura danych historycznych, metoda klasyfikacji obserwacji, wybór progu decyzyjnego lub instytucjonalne skutki wykorzystania predykcji. W efekcie system technicznie sprawny może uczestniczyć w procesie społecznie wadliwym, gdyż procedura matematyczna odpowiada na pytanie inne niż to, które powinno zostać rozstrzygnięte z perspektywy praw i potrzeb człowieka. Rozbieżność ta ma charakter epistemologiczny: każdy model jest uproszczoną reprezentacją rzeczywistości, a jego konstrukcja wymaga decyzji o tym, co uznamy za istotne, co da się zmierzyć i jakie zależności posłużą do prognozowania przyszłości. W uczeniu maszynowym zmienne wejściowe niosą wybrane informacje o obiekcie, zmienna docelowa określa rezultat do przewidzenia, a funkcja straty definiuje koszt błędów w przyjętej formalizacji. Decyzje te nie są neutralne poznawczo, ponieważ przesądzają o tym, jaka część świata społecznego uzyska reprezentację matematyczną. System oceniający kandydatów do pracy może analizować historię zatrudnienia czy wykształcenie, lecz niekoniecznie odróżni rzeczywiste kompetencje od przywilejów wynikających z pochodzenia społecznego.

Podobnie model oceniający ryzyko zdrowotne może opierać się na historii wydatków medycznych, choć koszty leczenia zależą nie tylko od stanu zdrowia, ale i od zasobów finansowych pacjenta czy dostępności usług. Z kolei model przewidujący ryzyko przestępczości często analizuje wcześniejsze zatrzymania, które nie są bezpośrednią ani pełną miarą wszystkich popełnionych czynów zabronionych. To właśnie w ten sposób działa Matematyka uprzedzeń i cicha produkcja nierówności.

Rozdźwięk między precyzją pomiaru a trafnością konstruktu

Problem ten można wyjaśnić poprzez rozróżnienie między trafnością konstruktu teoretycznego a poprawnością jego operacjonalizacji. W naukach społecznych pojęcia takie jak kompetencja, dobrostan, ryzyko, skłonność do współpracy, jakość życia czy sprawiedliwość nie są zazwyczaj bezpośrednio obserwowalnymi wielkościami fizycznymi. Wymagają one stworzenia wskaźników, których zdolność do reprezentowania badanego zjawiska powinna zostać poddana niezależnej ocenie. W psychometrii i metodologii badań empirycznych szczególną wagę przywiązuje się zatem do trafności pomiaru, odróżniając ją od jego rzetelności. Pomiar może być bowiem powtarzalny i wewnętrznie spójny, a mimo to systematycznie nie obejmować rzeczywistej właściwości, którą deklaruje mierzyć. Algorytm dostarczający konsekwentnie błędnej reprezentacji potrzeb społecznych nie staje się sprawiedliwy dzięki temu, że jego wyniki są stabilne. Podobnie prognoza

może być statystycznie dobrze dopasowana do historycznych obserwacji, a zarazem opierać się na danych odzwierciedlających wcześniejsze niesprawiedliwości.

Właśnie w tej przestrzeni między konstruktem a jego wskaźnikiem rozwija się znaczna część problemów określanych zbiorczym terminem stronniczości algorytmicznej. Harini Suresh i John Guttag, analizując źródła szkód na kolejnych etapach życia modeli uczenia maszynowego, wyróżnili między innymi stronniczość historyczną, reprezentacyjną, pomiarową, agregacyjną, związaną z uczeniem, ewaluacyjną oraz wdrożeniową. Taka klasyfikacja jest znacznie użyteczniejsza niż ogólny postulat eliminowania uprzedzeń z danych, ponieważ pozwala rozpoznać odmienne mechanizmy powstawania szkody i powiązać je z odpowiednimi metodami interwencji.

Stronniczość historyczna może występować nawet przy wiernym odwzorowaniu istniejącej rzeczywistości, jeżeli sama ona zawiera utrwalone nierówności. Stronniczość reprezentacyjna pojawia się wtedy, gdy określone populacje są niedostatecznie obecne w zbiorze obserwacji. Stronniczość pomiarowa dotyczy nieadekwatnych wskaźników, natomiast agregacyjna wynika z niewłaściwego traktowania zróżnicowanych populacji jako jednorodnej całości. Problemy ewaluacyjne ujawniają się, gdy testy nie odpowiadają rzeczywistym warunkom użytkowania, zaś wdrożeniowe powstają wtedy, gdy poprawnie funkcjonujący model zostaje użyty w niewłaściwej instytucji, wobec innej populacji albo do celu, którego nie uwzględniono podczas projektowania.

W tym kontekście należy wysnuć istotny wniosek metodologiczny: naprawa stronniczego systemu nie może automatycznie sprowadzać się do dodania danych albo zwiększenia liczby parametrów modelu. Jeżeli nieprawidłowo określono zjawisko będące przedmiotem prognozy, większa ilość danych może jedynie zwiększyć precyzję przewidywania niewłaściwej wielkości. Jeżeli źródłem problemu jest historyczna nierówność dostępu do określonego dobra, wierniejsze odtwarzanie tej historii może utrwalać jego niesprawiedliwą dystrybucję.

Jeżeli natomiast szkoda powstaje dopiero podczas wykorzystywania predykcji przez instytucję, nawet doskonała techniczna korekta modelu może okazać się niewystarczająca bez zmiany praktyki organizacyjnej. Istota problemu wymaga zatem odróżnienia trzech pytań: czy model poprawnie reprezentuje wybrane dane, czy wybrane dane odpowiednio reprezentują badane zjawisko oraz czy zastosowanie uzyskanego wyniku jest uzasadnione społecznie, moralnie i prawnie. Każde z tych pytań należy do innego poziomu analizy i wymaga odrębnego aparatu badawczego.

Jednym z najłatwiejszych do formalnego uchwycenia, a zarazem społecznie doniosłych mechanizmów zniekształcania wyników uczenia maszynowego jest nierównowaga klasowa, określana w literaturze jako **class imbalance**. W klasyfikacji nadzorowanej problem ten pojawia się wtedy, gdy obserwacje należące do różnych kategorii występują w zbiorze danych z bardzo

odmienną częstością. Sama nierównowaga nie stanowi automatycznie błędu metodologicznego, ponieważ wiele rzeczywistych zjawisk charakteryzuje się nierównomiernym rozkładem. Rzadkie choroby, niektóre przestępstwa, awarie infrastruktury krytycznej czy określone zdarzenia finansowe występują znacznie rzadziej niż ich brak. Niebezpieczeństwo powstaje wówczas, gdy przyjęta funkcja celu i sposób oceny modelu umożliwiają osiąganie wysokiego wyniku ogólnego kosztem niemal całkowitego pomijania zdarzeń rzadkich. Model może zostać uznany za skuteczny, ponieważ poprawnie klasyfikuje przeważającą część obserwacji, choć zawodzi dokładnie tam, gdzie jego zastosowanie miałoby największą wartość społeczną.

Wysoka dokładność modelu nie gwarantuje sprawiedliwości

Aby zobrazować ten mechanizm, posłużmy się przykładem hipotetycznego systemu medycznego, którego zadaniem jest wykrywanie choroby występującej u jednego procenta populacji. Jeśli analizie poddamy dziesięć tysięcy osób, przeciętnie chorych będzie około stu, podczas gdy pozostałe dziewięć tysięcy dziewięćset osób pozostanie zdrowych. Model, który każdą osobę zakwalifikuje jako zdrową, osiągnie ogólną dokładność na poziomie dziewięćdziesięciu dziewięciu procent.

Z perspektywy medycznej jego wartość diagnostyczna w zakresie wykrywania choroby pozostanie jednak zerowa, gdyż nie rozpozna on ani jednego rzeczywistego przypadku. Wysoka dokładność będzie tu jedynie rezultatem struktury populacji, a nie zdolności systemu do realizacji celu klinicznego. Ten eksperyment matematyczny dowodzi, że wskaźnik poprawnych klasyfikacji może być informacyjnie nieprzydatny, jeżeli nie zostanie zestawiony z częstością zdarzenia, rodzajami błędów i konsekwencjami ich wystąpienia. Problem pojawia się również w przypadku znacznie bardziej zaawansowanych modeli, które mimo korzystnych parametrów diagnostycznych mogą generować problematyczne wyniki.

Przyjmijmy, że system posiada dziewięćdziesięcioprocentową czułość i pięcioprocentowy odsetek wyników fałszywie dodatnich wśród osób zdrowych przy zachowaniu częstości występowania choroby na poziomie jednego procenta. Przy badaniu dziesięciu tysięcy osób model prawidłowo rozpozna około dziewięćdziesięciu przypadków, ale jednocześnie błędnie wskaże chorobę u około czterystu dziewięćdziesięciu pięciu osób zdrowych. W efekcie spośród wszystkich dodatnich wskazań jedynie około piętnaście procent będzie odpowiadało rzeczywistym przypadkom choroby. Zjawisko to wynika z twierdzenia Bayesa i pokazuje zasadnicze znaczenie częstości bazowej zdarzenia (**base rate**) dla interpretowania wyników testów. System o pozornie bardzo dobrych

parametrach może generować znaczną liczbę fałszywych alarmów, jeśli jest wykorzystywany wobec populacji, w której badane zdarzenie występuje rzadko.

Nie oznacza to automatycznie braku użyteczności badania przesiewowego – jego wartość może polegać przecież na kierowaniu pacjentów do dalszej diagnostyki. Dowodzi to jednak, że wynik modelu musi być interpretowany w kontekście częstości choroby, kosztów błędów i dostępności procedur potwierdzających. Formalny aparat statystyczny pozwala w tym miejscu rozróżnić kilka wskaźników, które w potocznej dyskusji bywają niesłusznie utożsamiane.

Czułość określa odsetek rzeczywistych przypadków pozytywnych prawidłowo rozpoznanych przez system, natomiast swoistość odnosi się do odsetka przypadków negatywnych poprawnie sklasyfikowanych jako negatywne. Precyzja, określana również jako dodatnia wartość predykcyjna, pokazuje, jaka część pozytywnych wskazań rzeczywiście odpowiada zdarzeniu. Zrównoważona dokładness uwzględnia jednocześnie czułość i swoistość, ograniczając wpływ dominacji jednej klasy na zagregowaną ocenę. Miara F1 łączy precyzję i czułość za pomocą średniej harmonicznej, lecz nie uwzględnia wprost prawdziwie negatywnych wyników.

Pole pod krzywą ROC dostarcza informacji o zdolności modelu do porządkowania przypadków pozytywnych i negatywnych, ale nie przesądza o użyteczności konkretnego progu decyzyjnego. W zadaniach obejmujących rzadkie zdarzenia szczególnie przydatne bywa analizowanie krzywej precyzja-czułość, ponieważ uwidacznia ona ograniczenia, które nie zawsze są czytelne przy wykorzystaniu samego ROC-AUC. Wybór metryki pozostaje nierozerwalnie związany z funkcją społeczną systemu.

W diagnostyce chorób o wysokiej śmiertelności kluczowe jest ograniczanie liczby wyników fałszywie negatywnych, podczas gdy w systemach podejmujących decyzje o ingerencji w wolność osobistą nieproporcjonalnie wysokie znaczenie może mieć koszt błędnego przypisania zagrożenia. W zastosowaniach finansowych istotne są zarówno straty wynikające z niespłaconych zobowiązań, jak i koszty nieuzasadnionego wykluczania wiarygodnych klientów. Żadna uniwersalna miara nie zastąpi więc analizy konsekwencji decyzji. Należy bowiem rozpoznać, że błędy statystyczne mogą generować szkody nieodwracalne lub trudne do naprawienia.

Odrzucenie kandydata w procesie rekrutacji może pozbawić go możliwości zdobycia doświadczenia zawodowego, co następnie wpłynie na jego przyszłą ocenę przez kolejne algorytmy. Tak powstaje relacja między błędem pojedynczego modelu a długookresowym procesem kumulowania nierówności.

Jeszcze subtelniejszym mechanizmem zniekształcania interpretacji statystycznych jest paradoks Simpsona, którego współczesne omówienie wiąże się z pracą Edwarda H. Simpsona z 1951 roku,

choć pokrewne własności rozkładów warunkowych analizowano wcześniej. Paradoks ten polega na możliwości wystąpienia odmiennej, a w szczególnych przypadkach przeciwnej zależności między zmiennymi w danych zagregowanych niż w odpowiednio wyodrębnionych podgrupach. Nie oznacza to sprzeczności matematycznej, ponieważ wskaźniki zagregowane są średnimi ważonymi, a struktura wag może różnić się między porównywanymi populacjami. Zjawisko to staje się szczególnie istotne wtedy, gdy badacz interpretuje ogólną zależność jako dowód mechanizmu przyczynowego, nie uwzględniając zmiennych wpływających jednocześnie na strukturę populacji i prawdopodobieństwo badanego rezultatu.

Paradoks Simpsona a błędna interpretacja danych

Klasycznego materiału empirycznego dostarcza analiza rekrutacji na studia podyplomowe Uniwersytetu Kalifornijskiego w Berkeley z 1973 roku, opublikowana przez Petera J. Bickela, Eugene'a A. Hammela i J. Williama O'Connella na łamach czasopisma **Science** w 1975 roku. W zestawieniu obejmującym całą uczelnię znalazło się 12 763 podań, z czego 8442 złożyły mężczyzny, a 4321 kobiety.

Ogólny odsetek przyjęć wynosił około 44% wśród mężczyzn i 35% wśród kobiet, co na poziomie agregatu sugerowało znaczącą nierówność. Jednak po uwzględnieniu specyfiki poszczególnych wydziałów okazało się, że kobiety częściej aplikowały do jednostek o wyższej konkurencji i niższej stopie przyjęć, podczas gdy mężczyźni relatywnie częściej ubiegali się o miejsca na wydziałach o korzystniejszych wskaźnikach. Oznaczało to, że obserwowana dysproporcja nie mogła zostać w prosty sposób przypisana dyskryminacyjnym decyzjom komisji rekrutacyjnych. Szczegółowa interpretacja wymaga ponadto rozróżnienia pełnego zbioru danych od jego najbardziej rozpowszechnionej dydaktycznie części. Popularny zbiór **UCBAdmissions**, dostępny między innymi w środowisku R, obejmuje 4526 zgłoszeń do sześciu największych wydziałów. W tej ograniczonej populacji przyjęto 1198 spośród 2691 mężczyzn (ok. 44,5%) oraz 557 spośród 1835 kobiet (ok. 30,4%).

Dane te nie są sprzeczne z ogólnouniwersyteckim zestawieniem, gdyż dotyczą innego zakresu obserwacji. Analiza sześciu wydziałów pokazuje, że w czterech z nich odsetek przyjmowanych kobiet był wyższy niż mężczyzn, mimo odwrotnej zależności na poziomie zagregowanym. Nie oznacza to jednak, że każdy wydział faworyzował kobiety ani że pierwotna analiza automatycznie wykluczyła wszelkie przejawy nierówności społecznych. Badacze wskazywali, że kluczowe znaczenie mogą mieć wcześniejsze etapy edukacyjnej selekcji, wpływające na wybór kierunków studiów, poziom przygotowania akademickiego i strukturę aspiracji zawodowych. Paradoks Simpsona nie

unieważnia zatem problemu dyskryminacji, lecz wymaga precyzyjnego określenia miejsca, w którym ewentualna niesprawiedliwość powstaje.

Formalne wyjaśnienie tego mechanizmu wynika z prawa całkowitego prawdopodobieństwa. Ogólne prawdopodobieństwo zdarzenia w danej grupie jest sumą prawdopodobieństw warunkowych występujących w poszczególnych podgrupach, ważonych ich udziałem w całej populacji. Jeżeli struktura uczestnictwa w podgrupach różni się między porównywanymi kategoriami, to wskaźniki zagregowane mogą pozostawać odmienne nawet przy bardzo zbliżonych prawdopodobieństwach warunkowych. W praktyce oznacza to, że porównywanie dwóch zbiorowości bez kontroli ich składu może prowadzić do błędnego przypisywania wpływu określonej cesze, podczas gdy obserwowany rezultat wynika przede wszystkim z odmiennych warunków uczestnictwa w procesie.

Zjawisko to znajduje zastosowanie w analizach medycznych, gospodarczych, edukacyjnych, demograficznych i politycznych. Może pojawiać się przy porównywaniu wyników szpitali obsługujących pacjentów o różnym stopniu zaawansowania chorób, szkół o odmiennym profilu uczniów czy przedsiębiorstw funkcjonujących w różnych segmentach rynku.

Błąd selekcji i pułapka danych kontrfaktycznych

Interpretacja paradoksu Simpsona prowadzi jednak do problemu znacznie głębszego niż samo rozdzielanie danych na podgrupy. Judea Pearl, rozwijając teorię wnioskowania przyczynowego, wykazał, jak istotne jest formalne odróżnienie zależności obserwacyjnych od skutków interwencji. Statystyczna korelacja między dwiema zmiennymi nie musi bowiem oznaczać, że zmiana jednej z nich wywoła zmianę drugiej. Badacz powinien najpierw ustalić, czy dodatkowa zmienna jest wspólną przyczyną analizowanych zjawisk, mediatorem określonego oddziaływania, czy też wspólnym skutkiem kilku procesów. W pierwszym przypadku uwzględnienie tej zmiennej może ograniczyć zniekształcenie wyniku; w drugim – prowadzić do pominięcia części rzeczywistego oddziaływania, natomiast w trzecim nieostrożne warunkowanie może samo w sobie wygenerować pozorną zależność. Dlatego popularny postulat uwzględniania wszystkich dostępnych zmiennych nie jest poprawną, uniwersalną strategią analizy przyczynowej. Szczególnie interesującą ilustrację ostatniego problemu stanowi paradoks Berksona, związany z tak zwanym błędem selekcji przez wspólny skutek (**collider bias**).

Wyobraźmy sobie proces rekrutacyjny, w którym przyjęcie kandydata zależy zarówno od kompetencji intelektualnych, jak i doświadczenia praktycznego. Jeżeli badacz analizuje wyłącznie osoby już zatrudnione, może zaobserwować pozorną ujemną zależność między tymi cechami, nawet jeśli w całej populacji kandydatów są one całkowicie niezależne. Wynika to z faktu, że osoby z

bardzo wysokim poziomem jednej cechy mogły zostać przyjęte mimo niższej wartości drugiej. Selekcja według wspólnego rezultatu tworzy zatem zależność, która nie występuje w populacji wyjściowej. W systemach AI problem ten pojawia się podczas uczenia modeli na danych pochodzących wyłącznie od osób wcześniej zaakceptowanych przez instytucję. Bank dysponuje szczegółowymi danymi o spłacie kredytu przede wszystkim w przypadku klientów, którym finansowanie faktycznie przyznano; pracodawca zna rzeczywiste efekty pracy głównie osób zatrudnionych, a uczelnia obserwuje przebieg studiów tych, którzy zostali przyjęci. Brak danych o konsekwencjach alternatywnych decyzji uniemożliwia wiarygodne oszacowanie, jak zachowałyby się osoby wcześniej wykluczone. Problem ten wiąże się bezpośrednio z zagadnieniem brakujących wyników kontrfaktycznych.

W klasycznym modelu przyczynowym Donalda Rubina oraz rozwijanych później metodach analizy efektów oddziaływania rozróżnia się możliwe rezultaty, które wystąpiłyby w odmiennych warunkach. Nie można jednak bezpośrednio zaobserwować wszystkich alternatywnych historii tej samej jednostki. Kredytobiorca, któremu odmówiono finansowania, nie dostarcza informacji o tym, czy spłaciłby nieudzielony kredyt. Kandydat odrzucony w rekrutacji nie może wykazać wyników pracy w organizacji, która go nie zatrudniła, a osoba niezakwalifikowana do programu medycznego nie ujawnia rezultatów opieki, której nie otrzymała.

System uczący się z historycznych decyzji może zatem utrwalać wcześniejsze kryteria selekcji nie dlatego, że zostały one empirycznie potwierdzone jako optymalne, lecz dlatego, że alternatywne rezultaty pozostają nieobserwowalne. Powstaje w ten sposób niebezpieczna asymetria poznawcza: decyzje podejmowane w przeszłości określają, jakie dowody będą dostępne przy ocenie przyszłych decyzji.

Wzajemna sprzeczność matematycznych kryteriów sprawiedliwości

Matematyczne badania nad sprawiedliwością algorytmiczną doprowadziły do opracowania wielu odmiennych kryteriów, których wybór zależy od charakteru problemu i przyjętej koncepcji równego traktowania. Jednym z najbardziej intuicyjnych jest parytet demograficzny (**demographic parity**), wymagający, aby odsetek pozytywnych decyzji był jednakowy w porównywanych grupach. W rekrutacji oznaczałoby to równy udział kandydatów zakwalifikowanych do kolejnego etapu, a w procesie kredytowym – podobny odsetek przyznanych finansowań. Kryterium to sprawdza się zwłaszcza wtedy, gdy celem jest ograniczanie wykluczenia dostępowego, jednak nie rozstrzyga ono automatycznie o sprawiedliwości indywidualnej. Jeśli populacje różnią się cechami merytorycznie i

prawnie uzasadnionymi, wymuszanie identycznych proporcji może kolidować z innymi celami procesu decyzyjnego. Z drugiej strony, pozornie obiektywne różnice mogą być jedynie wynikiem wcześniejszych nierówności, więc samo stwierdzenie odmiennej struktury populacji nie wystarczy, by uzasadnić zróżnicowane traktowanie.

Alternatywną koncepcję stanowi równość szans (**equality of opportunity**) w ujęciu formalnym zaproponowanym przez Moritza Hardta, Erica Price'a i Nathana Srebro. Według tej miary sprawiedliwości odsetek prawidłowo rozpoznanych przypadków pozytywnych powinien być porównywalny we wszystkich grupach. Rozszerzeniem tego podejścia jest warunek **equalized odds**, który wymaga jednoczesnego wyrównania wskaźników prawdziwie pozytywnych oraz fałszywie pozytywnych rozpoznań. Istotą tej metody jest kontrola nad tym, jak konkretne rodzaje błędów rozkładają się między populacjami. W systemach oceny ryzyka przestępczości wymóg ten oznacza dążenie do sytuacji, w której jedna grupa nie ponosi nieproporcjonalnie wysokich koszty błędnych alarmów, podczas gdy inna częściej uzyskuje nieuzasadnione kwalifikacje do kategorii niskiego ryzyka. Kryterium to ma jednak swoje ograniczenia: opiera się na założeniu, że etykiety opisujące rzeczywisty rezultat są wysokiej jakości. Jeśli zmienna docelowa jest zniekształcona, matematyczne wyrównanie błędów nie musi przynieść sprawiedliwych konsekwencji społecznych.

Kolejnym kryterium jest kalibracja predykcji (**calibration**), czyli zgodność między przypisywanym prawdopodobieństwem a rzeczywistą częstością zdarzenia. Jeśli system przypisuje danej grupie trzydziestoprocentowe ryzyko, to w dobrze skalibrowanym modelu około trzydziestu procent porównywalnych przypadków powinno zrealizować przewidywane zdarzenie. Kalibrację można oceniać oddzielnie dla różnych grup, co jest kluczowe w procedurach wysokiego ryzyka. Nie należy jej jednak utożsamiać z równością wskaźników błędnych decyzji po zastosowaniu konkretnego progu klasyfikacyjnego. Dwa modele mogą być poprawnie skalibrowane, a jednocześnie generować różne proporcje fałszywych alarmów, o ile w analizowanych grupach różni się częstość zdarzenia oraz rozkład wyników predykcyjnych. Jon Kleinberg, Sendhil Mullainathan i Manish Raghavan wykazali, że określone intuicyjnie atrakcyjne warunki sprawiedliwości probabilistycznych ocen ryzyka pozostają wzajemnie niezgodne poza wyjątkowymi sytuacjami. Niezależnie Alexandra Chouldechova przeanalizowała podobne ograniczenia na przykładzie narzędzi prognozowania recydywy. W przypadku populacji różniących się częstością badanego zdarzenia nie da się jednocześnie zapewnić kalibracji, identycznego rozkładu błędów i jednakowej interpretacji wyników.

Matematyczne konflikty metryk wymuszają wybory normatywne

Twierdzenia te mają ścisły charakter matematyczny, jednak ich interpretacja normatywna wymaga ostrożności. Nie dowodzą one, że sprawiedliwość algorytmiczna jest niemożliwa lub że wybór kryteriów zawsze pozostaje arbitralny. Wykazują natomiast, że sprawiedliwość nie jest jedną właściwością modelu, którą można bez konfliktów maksymalizować za pomocą prostego parametru. Różne koncepcje równości mogą bowiem wymagać odmiennych sposobów rozkładania kosztów błędów i korzyści wynikających z podejmowanych decyzji.

Zależność tę można zobrazować przykładem hipotetycznego systemu klasyfikacyjnego, w którym dwie grupy osób różnią się częstością występowania danego zdarzenia – w pierwszej grupie dotyczy ono 10% populacji, a w drugiej 30%. Przyjmijmy, że model osiąga w obu grupach identyczną czułość (80%) oraz taki sam odsetek fałszywych alarmów (10%). W pierwszej grupie system prawidłowo wskaże średnio ośmiu przypadków pozytywnych, ale błędnie zakwalifikuje dziewięć osób negatywnych; dodatnia wartość predykcyjna wyniesie zatem około 47%. W drugiej grupie poprawnie rozpoznanych zostanie przeciętnie dwudziestu cztery przypadki pozytywne przy siedmiu fałszywie dodatnich, co daje dodatnią wartość predykcyjną na poziomie około 77%.

Identyczne parametry czułości i fałszywych alarmów nie gwarantują więc równej precyzji predykcji, jeśli różni się częstość bazowa zdarzenia. Jest to cecha rachunku prawdopodobieństwa, a nie dowód na to, że któraś z grup powinna być traktowana korzystniej. Matematyczne konflikty sprawiedliwości wymuszają zatem określenie celów normatywnych, których nie da się wyprowadzić wyłącznie z danych. W teorii Johna Rawlsa kluczowe jest tu znaczenie zasady równej podstawowej wolności oraz koncepcji sprawiedliwej równości szans, wymagającej analizy społecznych warunków uczestnictwa w instytucjach. Amartya Sen wskazywał natomiast na ograniczenia postrzegania sprawiedliwości jedynie przez pryzmat formalnych zasobów i wyników, proponując analizę rzeczywistych możliwości działania człowieka – perspektywę tę rozwijała Martha Nussbaum, wskazując na podstawowe zdolności konieczne do godnego życia.

Przeniesienie tych koncepcji na grunt AI nie polega na prostym utożsamieniu metryki matematycznej z teorią filozoficzną. Pozwala raczej zrozumieć, że równy odsetek pozytywnych decyzji nie musi oznaczać równej dostępności rzeczywistych możliwości, podobnie jak identyczne prawdopodobieństwo błędnej klasyfikacji nie musi prowadzić do jednakowych konsekwencji dla osób o różnym statusie społecznym i ekonomicznym. Jednym z najbardziej przekonujących empirycznie przykładów nieadekwatnej operacjonalizacji celu jest badanie Ziada Obermeyera, Briana Powersa, Christine Vogeli i Sendhila Mullainathana z 2019 roku, opublikowane w

czasopiśmie **Science** pod tytułem **Dissecting racial bias in an algorithm used to manage the health of populations**.

Błąd zmiennej zastępczej utrwala nierówności systemowe

Autorzy analizowali komercyjny algorytm służący do identyfikacji pacjentów wymagających intensywniejszego wsparcia medycznego. System opierał się na prognozie przyszłych kosztów opieki zdrowotnej, traktując je jako przybliżenie stopnia potrzeb zdrowotnych. Z perspektywy zarządzania świadczeniami założenie to mogło wydawać się racjonalne – osoby wymagające skomplikowanego leczenia zazwyczaj generują wyższe wydatki. Problem polegał jednak na tym, że koszty medyczne nie zależą wyłącznie od obciążenia chorobowego; są one uwarunkowane dostępnością usług, historią korzystania z opieki, strukturą ubezpieczenia, sytuacją społecznoekonomiczną oraz nierównościami instytucjonalnymi. Badacze wykazali, że przy podobnym wyniku oceny ryzyka pacjenci czarnoskórzy byli przeciętnie bardziej chorzy niż biali, co dowodziło nieadekwatności kosztów jako wspólnej miary rzeczywistych potrzeb zdrowotnych. Po zastąpieniu tego wskaźnika lepszą reprezentacją stanu zdrowia oszacowano, że udział pacjentów czarnoskórych kwalifikowanych do dodatkowego wsparcia mógłby wzrosnąć z 17,7% do 46,5%. Różnica ta nie wynikała z konieczności zwiększenia liczby parametrów modelu ani z prostego dopisania zakazu dyskryminacji.

Źródłem błędu był wybór zmiennej docelowej: system odpowiadał na pytanie o przewidywany koszt opieki, podczas gdy celem instytucjonalnym powinno być rozpoznawanie osób wymagających wsparcia zdrowotnego. W ten sposób matematyczna zgodność z funkcją celu stawała się mechanizmem systematycznego ograniczania dostępu części pacjentów do świadczeń.

Przypadek ten jest istotny dla filozofii i metodologii AI, gdyż ujawnia zjawisko stronniczości zmiennej zastępczej (**proxy bias**). Instytucja może nie wykorzystywać bezpośrednio cechy chronionej, a nawet deklarować chęć sprawiedliwego traktowania pacjentów. Jeśli jednak przyjęty wskaźnik jest historycznie powiązany z nierównym dostępem do usług, system będzie odtwarzał te nierówności przy przyszłej alokacji zasobów. Problem ma charakter przyczynowy: niższe wydatki mogą oznaczać lepszy stan zdrowia, ale mogą być również efektem barier w dostępie do leczenia.

Model, który nie rozróżnia tych mechanizmów, przypisuje tej samej obserwacji finansowej błędne znaczenie kliniczne. W praktyce instytucjonalnej źle zdefiniowany wskaźnik zaczyna zastępować rzeczywisty cel działalności, a jego precyzyjna matematyczna optymalizacja może utrudnić

zauważenie pomyłki. Zjawisko to wiąże się z prawem Goodharta, według którego miara wykorzystywana jako cel traci swoje pierwotne właściwości informacyjne. Gdy organizacja zaczyna nagradzać konkretny wskaźnik, uczestnicy mogą dostosowywać zachowanie do sposobu jego obliczania, a nie do dobra, które miał on reprezentować. Donald T. Campbell analizował podobne konsekwencje stosowania wskaźników ilościowych w ocenie procesów społecznych, wskazując na ryzyko ich deformacji pod wpływem presji organizacyjnej.

W edukacji koncentracja wyłącznie na wynikach testów może skłaniać do ograniczania programu nauczania do zadań egzaminacyjnych. W ochronie zdrowia premiowanie liczby obsłużonych pacjentów bywa przyczyną skracania konsultacji, a w administracji nacisk na szybkość rozpatrywania spraw może obniżać jakość uzasadnień decyzji. Zastosowanie AI zwiększa skalę tego ryzyka – system optymalizuje określoną miarę znacznie szybciej i konsekwentniej niż człowiek, nawet gdy przestaje ona odzwierciedlać rzeczywisty cel instytucji.

Stronniczość danych wynika z mechanizmów selekcji i pomiaru

Rozbudowanie matematycznego aparatu oceny sprawiedliwości wymaga uwzględnienia mechanizmów powstawania samych danych. W analizie społecznych zastosowań AI szczególne znaczenie zyskuje stronniczość selekcji (**selection bias**), która występuje, gdy prawdopodobieństwo znalezienia się obserwacji w zbiorze danych jest powiązane z właściwościami badanego zjawiska. Dane zgromadzone na platformie internetowej odzwierciedlają przede wszystkim zachowania jej użytkowników, a niekoniecznie całej populacji. Podobnie informacje pochodzące ze szpitali mogą pomijać osoby pozbawione dostępu do opieki medycznej.

Z kolei dane dotyczące kredytobiorców odnoszą się w znacznej mierze do osób, które przeszły już wcześniejszy proces selekcji bankowej. W każdym z tych przypadków zbiór treningowy stanowi produkt określonego mechanizmu instytucjonalnego, a nie losową próbę wszystkich potencjalnych użytkowników systemu. Traktowanie takiej próby jako reprezentatywnej może prowadzić do systematycznych błędów generalizacji.

Odrębny charakter ma stronniczość pomiarowa (**measurement bias**), powstająca wtedy, gdy wykorzystywane wskaźniki nie rejestrują badanego zjawiska w sposób porównywalny we wszystkich populacjach. Test kompetencji językowych może być trafny dla użytkowników określonego wariantu języka, lecz mniej dokładny dla osób posługujących się innymi dialektami. System analizy mowy może osiągać korzystne wyniki przy nagraniach studyjnych, ale wykazywać

wyższą stopę błędów w przypadku osób mówiących niewyraźnie, korzystających ze słabych mikrofonów lub komunikujących się w hałasie. Algorytm rozpoznawania obrazu może z kolei wykazywać różnice w jakości zależne od oświetlenia, modelu aparatu, odcienia skóry czy sposobu przygotowania danych.

W takich sytuacjach wyrównanie liczebności grup nie zawsze wystarcza, gdyż problem dotyczy samej jakości pomiaru. Konieczne staje się zatem badanie porównywalności konstruktu, niezawodności instrumentów i warunków rejestracji danych. Psychometria dostarcza tu kluczowego pojęcia niezmienniczości pomiarowej (**measurement invariance**). Jeśli model ma porównywać właściwości różnych grup, należy ocenić, czy te same wyniki rzeczywiście oznaczają podobny poziom badanego konstruktu. W klasycznych modelach pomiarowych rozróżnia się m.in. niezmienniczość strukturalną, metryczną i skalarną, które odnoszą się do kolejnych poziomów porównywalności narzędzia.

W zastosowaniach AI analogiczne zagadnienia pojawiają się przy ocenie kompetencji, stanów emocjonalnych, ryzyka czy zachowań społecznych. Nie można zakładać, że cecha wykorzystywana przez model posiada identyczne znaczenie w każdej kulturze, grupie zawodowej lub kategorii demograficznej. Szczególnie problematyczne jest przenoszenie modeli opracowanych w jednym środowisku językowym i kulturowym do populacji o odmiennych praktykach komunikacyjnych, normach społecznych i uwarunkowaniach instytucjonalnych.

Ostatnim mechanizmem jest stronniczość agregacyjna (**aggregation bias**). Występuje ona wtedy, gdy jedna funkcja predykcyjna zostaje zastosowana wobec populacji, w których relacje między cechami wejściowymi a przewidywanym rezultatem są istotnie odmienne. Przykładowo ten sam zestaw wskaźników klinicznych może mieć zupełnie inne znaczenie diagnostyczne dla pacjentów w różnym wieku lub osób z odmiennymi chorobami współistniejącymi.

Ryzyka uśredniania danych i niestabilność modeli w czasie

Jeden uniwersalny model może osiągać wysoki wynik ogólny, a jednocześnie systematycznie zawodzić w przypadku grup mniejszościowych. Z drugiej strony tworzenie odrębnych modeli dla bardzo małych populacji zwiększa ryzyko niestabilności estymacji i nadmiernego dopasowania, prowadząc do utraty informacji, którą można by zyskać dzięki wspólnemu uczeniu. Rozwiązania techniczne w tym zakresie obejmują modele hierarchiczne, uczenie wielozadaniowe, częściowe współdzielenie parametrów, ważenie obserwacji oraz metody dostrajania specyficzne dla danej

domeny. Wybór między modelem wspólnym a modelami grupowymi wymaga jednak analizy nie tylko jakości predykcji, ale i społecznej dopuszczalności stosowania zróżnicowanych procedur.

W tym obszarze istotną rolę odgrywają modele hierarchiczne rozwijane w statystyce bayesowskiej. Umożliwiają one jednoczesne wykorzystanie informacji wspólnej dla całej populacji przy jednoczesnym uwzględnieniu odmienności poszczególnych podgrup. Mechanizm częściowego łączenia informacji (**partial pooling**) pozwala ograniczać niestabilność estymacji w małych populacjach, nie narzucając im automatycznie identycznych parametrów. W praktyce może to mieć kluczowe znaczenie przy analizie lokalnych jednostek administracyjnych, niewielkich grup pacjentów lub rzadkich kategorii zdarzeń. Nie jest to jednak uniwersalne rozwiązanie problemu sprawiedliwości, gdyż modele hierarchiczne również zależą od poprawności założeń, właściwego doboru zmiennych i jakości danych. Ich zaletą jest raczej dostarczenie bardziej elastycznej reprezentacji zróżnicowania, które pozostaje niewidoczne przy mechanicznym stosowaniu jednej średniej dla całej populacji.

Przesunięcie rozkładu i niebezpieczeństwa przenoszenia modeli między środowiskami

Jednym z najważniejszych wyzwań współczesnego uczenia maszynowego jest zmiana właściwości danych pomiędzy etapem trenowania modelu a rzeczywistym środowiskiem jego zastosowania. Zjawisko to, określane zbiorczo jako przesunięcie rozkładu (**distribution shift**), obejmuje kilka odmiennych mechanizmów statystycznych. Przesunięcie kowariantów (**covariate shift**) występuje wtedy, gdy zmienia się rozkład cech wejściowych, podczas gdy zależność zmiennej docelowej od tych cech pozostaje stabilna. Przesunięcie etykiet (**label shift**) dotyczy zmian częstości klas przy określonych założeniach dotyczących rozkładu cech warunkowanych klasą. Zmiana koncepcji lub relacji predykcyjnej (**concept drift**) pojawia się natomiast wtedy, gdy modyfikuje się sama zależność między obserwowanymi właściwościami a rezultatem, który model ma przewidywać. Rozróżnienia te mają znaczenie praktyczne, ponieważ różne rodzaje przesunięcia wymagają odmiennych metod diagnostycznych i aktualizacyjnych.

W systemach finansowych zmiany sytuacji makroekonomicznej mogą modyfikować relacje między dochodem, zadłużeniem a prawdopodobieństwem niewypłacalności. W ochronie zdrowia pojawienie się nowych metod leczenia, zmiana struktury populacji lub rozwój epidemii mogą sprawić, że wcześniej wytrenowane modele staną się mniej skuteczne. W administracji publicznej nowelizacja przepisów może zmienić znaczenie danych opisujących sytuację obywatela, a w systemach rekomendacyjnych nowe praktyki kulturowe i zmiany w sposobie korzystania z platform prowadzą do stopniowego dezaktualizowania wcześniejszych wzorców zainteresowań. Model nie musi ulec technicznej awarii, aby utracić przydatność. Może nadal prawidłowo wykonywać

wszystkie operacje obliczeniowe, ale odnosić je do relacji statystycznych, które przestały odpowiadać rzeczywistości.

Geograficzna stronniczość danych i kolonialność cyfrowa

Kluczową kwestią jest stronniczość geograficzna, którą autor wskazuje jako konsekwencję nierównomiernej reprezentacji regionów świata w danych treningowych. Większość dużych zbiorów cyfrowych odzwierciedla przede wszystkim działalność społeczeństw z wysokim dostępem do internetu, rozbudowaną infrastrukturą dokumentacyjną i silną pozycją gospodarczą.

Rodzi to ryzyko traktowania określonych praktyk językowych, społecznych czy administracyjnych jako uniwersalnych, podczas gdy w rzeczywistości są one właściwe jedynie dla ograniczonej części populacji. Model trenowany głównie na danych anglojęzycznych może błędnie interpretować lokalne idiomy, specyficzne konstrukcje prawne, kontekst historyczny lub praktyki instytucjonalne innych państw. Z tego względu ocena jakości modeli wielojęzycznych nie powinna ograniczać się do sprawności tłumaczenia pojedynczych zdań, lecz obejmować semantyczną adekwatność odpowiedzi, rozumienie kontekstu oraz odporność na przenoszenie nieuprawnionych założeń między różnymi systemami społecznymi.

Problematykę tę można osadzić w szerszej debacie dotyczącej kolonialności danych i nierównego udziału społeczeństw w tworzeniu infrastruktury poznawczej. Badacze społeczeństwa cyfrowego zauważają, że choć dane do rozwoju modeli pochodzą od milionów użytkowników, możliwości ich przetwarzania, komercjalizacji i kontrolowania pozostają skoncentrowane w rękach nielicznych organizacji. Powstaje tym samym asymetria między podmiotami dostarczającymi zasobów informacyjnych a tymi, którzy posiadają zdolność do ich gospodarczej eksploatacji.

Nie oznacza to, że każda forma przetwarzania danych z różnych kręgów kulturowych jest niesprawiedliwa. Wymaga jednak świadomości warunków pozyskiwania informacji, praw użytkowników, reprezentatywności językowej oraz dostępności korzyści i lokalnej kontroli nad technologią. Stronniczość algorytmiczna może zatem być nie tylko problemem pomiarowym, lecz rezultatem globalnej struktury ekonomicznej, w której zdolność do wyznaczania standardów cyfrowych jest rozłożona nierównomiernie. Jednym z najbardziej niebezpiecznych mechanizmów reprodukowania tych nierówności jest sprzężenie zwrotne między działaniem modelu a danymi, które model wykorzystuje do dalszego uczenia. W klasycznym schemacie statystycznym często zakłada się istnienie zbioru obserwacji, na podstawie którego można przewidywać zdarzenia

zachodzące niezależnie od procesu prognozowania. W zastosowaniach społecznych założenie to jest jednak niewystarczające, ponieważ decyzja systemu sama wpływa na przyszłe zachowanie ludzi i prawdopodobieństwo wystąpienia określonych zdarzeń. Przykładowo algorytm rekomendacyjny decyduje o widoczności treści, kształtując liczbę interakcji, które następnie interpretuje jako miarę zainteresowania.

Sprzężenia zwrotne i pułapka samopotwierdzających się predykcji

System nadzorujący obszary miejskie, kierując większą uwagę służb do konkretnych lokalizacji, zwiększa prawdopodobieństwo wykrycia tam zdarzeń, co z kolei utwierdza system w pierwotnej ocenie ryzyka. Podobnie model rekrutacyjny może ograniczać dostęp do pracy osobom o określonym profilu; brak późniejszego doświadczenia zawodowego staje się wówczas kolejnym argumentem przeciwko ich zatrudnieniu.

Formalne badania nad tym zjawiskiem prowadzili m.in. Danielle Ensign, Sorelle Friedler, Scott Neville, Carlos Scheidegger oraz Suresh Venkatasubramanian. W pracy **Runaway Feedback Loops in Predictive Policing** z 2018 roku autorzy przedstawili model niebezpiecznych sprzężeń zwrotnych w predykcyjnym rozmieszczaniu patroli policyjnych. W określonych warunkach algorytm wielokrotnie kieruje zasoby do tych samych obszarów, gdyż intensyfikacja kontroli naturalnie prowadzi do większej liczby odnotowanych incydentów. System interpretuje te obserwacje jako potwierdzenie prognozy, choć różnica ta wynika z nierównomiernej intensywności wykrywania, a nie z rzeczywistego rozkładu zdarzeń. Badacze wykazali, że pomocne w diagnozowaniu problemu może być odróżnianie incydentów zgłaszanych niezależnie od policji od tych odkrywanych dzięki patrolom, choć nie zawsze eliminuje to błąd całkowicie. Mechanizm ten obrazuje, dlaczego brak refleksji nad genezą danych prowadzi do pozornie samopotwierdzających się predykcji.

Analogiczne procesy zachodzą w gospodarce platformowej. Systemy rekomendacyjne opierają się na sygnałach aktywności: kliknięciach, czasie oglądania czy interakcjach użytkowników. Jednak możliwość wygenerowania takich danych zależy od tego, czy treść została wyświetlona i gdzie w interfejsie ją umieszczono. Stronniczość prezentacji (**presentation bias**) może prowadzić do zawyżonej oceny obiektów o lepszej pozycji rankingowej, gdyż ich widoczność sama w sobie zwiększa szansę na interakcję. Z kolei stronniczość aktywności (**activity bias**) pojawia się, gdy niewielka grupa intensywnych użytkowników dostarcza nieproporcjonalnie dużo danych

treningowych. Model dostosowuje się wówczas do zachowań osób najbardziej aktywnych, marginalizując preferencje pozostałych.

Dopełnieniem tych zjawisk jest stronniczość świeżości (**recency bias**), która nadaje nadmierne znaczenie ostatnim zdarzeniom kosztem stabilnych wzorców długookresowych. Wszystkie te mechanizmy wynikają z konstrukcji funkcji celu i sposobu zbierania danych, a ich skutki potęguje regularne aktualizowanie modelu na podstawie jego własnych rekomendacji. W odpowiedzi na to rozwijane są metody ograniczania zniekształceń: kontrolowane eksperymenty, losowe próbkowanie ekspozycji czy modelowanie prawdopodobieństwa interakcji. Metody szacowania odwrotnością prawdopodobieństwa ekspozycji (**inverse propensity weighting**) pozwalają w pewnych warunkach ograniczyć błąd wynikający z nierównej widoczności obiektów. Nie są one jednak uniwersalne, gdyż wymagają poprawnego określenia mechanizmu selekcji i odpowiedniego pokrycia przestrzeni możliwych decyzji.

Eksperymentowanie z ekspozycją rodzi własne problemy etyczne, szczególnie w obszarze informacji zdrowotnych, politycznych lub treści oddziałujących na osoby wrażliwe. Projektowanie uczciwych systemów rekomendacyjnych wymaga zatem syntezy statystyki eksperymentalnej, ekonomii informacji, psychologii użytkownika i odpowiedzialności instytucjonalnej. Rozwój badań nad sprawiedliwością AI przyniósł liczne metody mitygacji nierówności na różnych etapach życia modelu. Na poziomie danych stosuje się m.in. ważenie obserwacji, zwiększanie reprezentacji grup niedostatecznie obecnych w próbie oraz syntetyczne generowanie przykładów. Klasycznym przykładem techniki tworzenia syntetycznych obserwacji dla klasy mniejszościowej jest algorytm SMOTE, przedstawiony przez Nitesha Chawlę i współautorów w 2002 roku.

Techniczne korekty nie zastąpią analizy źródeł niesprawiedliwości

W określonych zastosowaniach metoda ta może poprawić zdolność klasyfikatora do rozpoznawania rzadziej występujących przypadków, jednak jej skuteczność zależy od charakteru danych i geometrii przestrzeni cech. Mechaniczne generowanie dodatkowych obserwacji nie usuwa błędnych etykiet ani nie naprawia źle zdefiniowanego celu; może wręcz pogorszyć jakość modelu, jeśli syntetyczne przykłady nie oddają rzeczywistych zależności w populacji. Na etapie uczenia można modyfikować funkcję straty tak, by uwzględniała dodatkowe ograniczenia dotyczące rozkładu błędów, wrażliwości na konkretne przypadki lub kosztów błędnych decyzji. Autor książki wskazuje tutaj na wykorzystanie regularyzacji, uczenia adwersarialnego, ograniczeń sprawiedliwościowych oraz

optymalizacji wielokryterialnej. Przyjmowane kryterium sprawiedliwości musi być jednak uzasadnione w odniesieniu do konkretnego zastosowania.

Mechaniczne minimalizowanie różnicy między wskaźnikami dwóch grup może poprawić formalny wynik audytu, nie eliminując przy tym przyczyn historycznych nierówności. W niektórych przypadkach taka korekta zmienia relacje między kalibracją, czułością a precyzją, co wymusza nowe kompromisy statystyczne. Ocena skuteczności wymaga zatem analizy zarówno wybranej metryki sprawiedliwości, jak i realnego wpływu rozwiązania na rezultaty społeczne. Metody przetwarzania wyników po treningu (**post-processing**) pozwalają z kolei modyfikować progi decyzyjne, kalibrować wyniki lub dostosowywać klasyfikację do przyjętych ograniczeń. Rozwiązania te bywają użyteczne, gdy konstruktor nie ma pełnej kontroli nad architekturą modelu lub gdy instytucja chce poprawić właściwości istniejącego systemu.

Nie można jednak utożsamiać każdej korekty wyniku z powierzchownym ukrywaniem niesprawiedliwości. Jeśli zastosowanie odmiennego progu decyzyjnego jest rzetelnie uzasadnione i pozwala ograniczyć nieproporcjonalne koszty błędów, może ono stanowić element racjonalnej polityki zarządzania ryzykiem. Jednocześnie modyfikacja końcowego wyniku nie powinna zastępować analizy źródeł zniekształcenia – niektóre problemy wymagają bowiem zmiany danych, funkcji celu lub całego procesu instytucjonalnego. Autor książki przywołuje również koncepcję perturbacji różnorodności (**diversity perturbation**), rozumianą jako celowe przeciwdziałanie nadmiernej koncentracji rekomendacji i ograniczaniu ekspozycji mniej popularnych obiektów. W badaniach nad systemami informacyjnymi problem ten łączy się z optymalizacją wielokryterialną, która uwzględnia nie tylko trafność wyników, ale również ich różnorodność, nowość, pokrycie zbioru obiektów i sprawiedliwość wobec uczestników rynku. W praktyce może to oznaczać prezentowanie treści mniej podobnych do wcześniejszych wyborów użytkownika, by zwiększyć szansę na poznanie alternatywnych informacji. Sam wzrost różnorodności nie gwarantuje jednak wartości poznawczej ani sprawiedliwości; system musi wciąż rozpoznawać, czy proponowane treści są odpowiednie, wiarygodne i bezpieczne.

W zastosowaniach medycznych lub prawnych celowe obniżanie trafności informacji w imię abstrakcyjnej różnorodności mogłoby prowadzić do poważnych szkód.

Konflikty między prywatnością, wyjaśnialnością a statystyczną miarą sprawiedliwości

Zasadność takiej interwencji zależy zatem od charakteru domeny i rodzaju dobra, które system ma wspierać. Istotnym zagadnieniem pozostaje również relacja między ochroną prywatności a sprawiedliwością modeli. Techniki prywatności różnicowej (**differential privacy**) pozwalają ograniczyć wpływ pojedynczej obserwacji na wynik procedury statystycznej, co zmniejsza ryzyko ujawnienia informacji o uczestnikach zbioru danych. Mechanizmy te mogą jednak powodować spadek dokładności, szczególnie w przypadku niewielkich populacji i rzadkich kategorii obserwacji. Eugene Bagdasaryan, Omid Poursaeed i Vitaly Shmatikov wykazali w badaniach opublikowanych podczas NeurIPS 2019, że określone metody prywatnego uczenia sieci neuronowych mogą nieproporcjonalnie pogarszać skuteczność klasyfikacji wobec niedostatecznie reprezentowanych grup. Nie oznacza to, że ochrona prywatności jest zawsze sprzeczna ze sprawiedliwością; pokazuje natomiast, że realizacja jednego celu etycznego może generować koszty dla innego, a projektowanie odpowiedzialnych systemów wymaga analizy rzeczywistych konsekwencji przyjętych zabezpieczeń.

Odrębne znaczenie mają techniki wyjaśnialności: modele interpretowalne, analiza znaczenia cech, metody przybliżonego opisywania lokalnych decyzji oraz interpretacje kontrfaktyczne. Ich przydatność wynika z możliwości identyfikowania zależności, na których opiera się klasyfikacja, jednak nie wszystkie formy wyjaśnienia posiadają identyczną wartość epistemiczną. Metoda wskazująca zmienne, które najsilniej wpłynęły na wynik konkretnego modelu, nie musi bowiem dostarczać informacji o rzeczywistych przyczynach badanego zdarzenia. Lokalna interpretacja decyzji może być ponadto niestabilna, jeśli niewielka zmiana danych wejściowych prowadzi do istotnie odmiennego rozkładu znaczenia cech. W badaniach nad sprawiedliwością konieczne jest więc rozróżnienie wyjaśnienia działania modelu od wyjaśnienia społecznej przyczynowości oraz od uzasadnienia normatywnego decyzji. Są to trzy odmienne zadania, których nie da się automatycznie rozwiązać jednym narzędziem informatycznym.

Szczególne zagrożenie dla oceny systemów AI stanowi nadmierne przywiązanie do pojedynczej wartości wskaźnika jakości. Wyniki testów zależą bowiem od liczebności próby, sposobu jej doboru, reprezentacji przypadków trudnych, definicji prawidłowego rezultatu i metod agregowania obserwacji. Różnica między dwoma systemami może być statystycznie nieistotna albo ograniczona do konkretnego zakresu zastosowań, mimo że komunikacja marketingowa przedstawia ją jako przełom. Przy ocenie sprawiedliwości kluczowa jest niepewność związana z małymi podgrupami: model może osiągać pozornie doskonały wynik dla grupy obejmującej niewielką liczbę obserwacji, lecz oszacowanie to może być obarczone szerokim przedziałem ufności.

Dlatego audyty powinny obejmować odpowiednie przedziały niepewności, analizę wrażliwości na dobór próby, procedury walidacji oraz testy w rzeczywistych warunkach użytkowania. W statystyce klasycznej istotną rolę odgrywa rozróżnienie pomiędzy istotnością statystyczną a znaczeniem praktycznym. Przy bardzo dużych zbiorach danych nawet niewielkie różnice między grupami mogą wykazywać wysoką istotność statystyczną, choć ich rzeczywisty wpływ na dobrostan pozostaje ograniczony. Z drugiej strony w przypadku niewielkich populacji poważne nierówności mogą nie osiągać przyjętego progu istotności z powodu niedostatecznej mocy testu. Ocena powinna więc uwzględniać wielkość efektu, jego niepewność, kierunek potencjalnego oddziaływania oraz charakter zagrożonych dóbr. W zastosowaniach wysokiego ryzyka niewielka różnica procentowa może mieć ogromne znaczenie, jeśli dotyczy setek tysięcy decyzji albo zdarzeń o nieodwracalnych skutkach. Jednocześnie mechaniczne żądanie identycznych wskaźników dla wszystkich możliwych podgrup może prowadzić do niestabilnych i trudnych do interpretacji rezultatów.

Epistemiczna pokora i prymat norm prawnych nad optymalizacją

Problematyka ta prowadzi do istotnej koncepcji epistemicznej pokory systemów AI. Model nie powinien jedynie generować ocen i rekomendacji, lecz tam, gdzie jest to technicznie uzasadnione, dostarczać również informacji o ograniczeniach swojego zastosowania, stopniu niepewności oraz sytuacjach wykraczających poza zweryfikowany zakres kompetencji. W badaniach nad selektywną predykcją (**selective prediction**) rozwijane są rozwiązania pozwalające systemowi wstrzymać automatyczne rozstrzygnięcie i przekazać sprawę do dodatkowej oceny, gdy poziom pewności lub ryzyka nie spełnia przyjętych kryteriów. Metody konformalnej predykcji (**conformal prediction**) umożliwiają w określonych warunkach konstruowanie zbiorów predykcyjnych z formalnymi gwarancjami pokrycia, choć ich skuteczność zależy od założeń dotyczących danych, a gwarancje zagregowane nie zawsze przekładają się na jednakową ochronę każdej podgrupy.

Narzędzia te nie zastępują sprawiedliwości instytucjonalnej, lecz mogą ograniczać sytuacje, w których system przedstawia wynik jako definitywny mimo niewystarczających podstaw poznawczych. Z perspektywy filozofii odpowiedzialności kluczowe staje się rozróżnienie między niepewnością poznawczą a niepewnością wynikającą z losowości badanego procesu. Pierwsza może zmniejszać się wraz ze wzrostem wiedzy, poprawą jakości danych i doskonaleniem modelu; druga natomiast pozostaje związana z rzeczywistą zmiennością zdarzeń. W praktyce oba rodzaje niepewności często współwystępują, a ich precyzyjne oddzielenie bywa trudne.

Model może wykazywać znaczną pewność numeryczną, mimo że został zastosowany wobec populacji odmiennej od treningowej. Może również przypisywać bardzo precyzyjne prawdopodobieństwo zdarzeniu, którego rzeczywistych determinant nie rozpoznano. Odpowiedzialne wykorzystanie predykcji wymaga zatem oceny nie tylko poziomu deklarowanej pewności, lecz także wiarygodności mechanizmu jej wyznaczania. Szczególnie w obszarach oddziałujących na prawa jednostki powinna istnieć możliwość odmowy automatycznego rozstrzygnięcia i uruchomienia procedury bardziej adekwatnej do charakteru problemu. Zasadniczym rezultatem interdyscyplinarnej analizy stronniczości jest rozpoznanie, że każda funkcja celu stanowi formalizację określonej koncepcji wartości. Nawet gdy system został zaprojektowany wyłącznie w celu minimalizowania błędu predykcji, konieczne było wcześniejsze określenie, jakie zdarzenie ma zostać przewidziane, które błędy uznaje się za istotne oraz jaką wagę przypisuje się poszczególnym konsekwencjom. W zastosowaniach społecznych decyzje te nie mogą być utożsamiane wyłącznie z technicznym doborem hiperparametrów; wyrażają one bowiem określony sposób organizowania relacji między interesem jednostki, efektywnością instytucji, bezpieczeństwem zbiorowym a społeczną dystrybucją ryzyka.

Odpowiedzialność za sztuczną inteligencję rozpoczyna się zatem wcześniej niż programowanie i trenowanie modelu. Obejmuje ona już etap definiowania problemu oraz uzasadniania, dlaczego właśnie to zagadnienie powinno zostać rozwiązane za pomocą technologii obliczeniowej. Filozofia prawa dostarcza tutaj ważnych narzędzi analitycznych. Ronald Dworkin zwracał uwagę na znaczenie zasad moralnych w interpretacji prawa oraz na ograniczenia mechanicznego stosowania reguł oderwanych od ich normatywnego uzasadnienia. Robert Alexy rozwijał teorię zasad jako nakazów optymalizacyjnych, których realizacja wymaga uwzględniania możliwości faktycznych i prawnych, w tym kolizji z innymi zasadami. Nie oznacza to jednak, że matematyczna optymalizacja modelu AI jest odpowiednikiem prawnego ważenia zasad. Wręcz przeciwnie, zestawienie obu koncepcji pozwala dostrzec zasadniczą różnicę między obliczeniowym minimalizowaniem funkcji straty a prawnym uzasadnianiem ograniczeń praw podstawowych. W systemie demokratycznym decyzje dotyczące wolności, równości, prywatności i dostępu do dóbr powinny być poddawane kontroli według właściwych norm prawnych, a nie wyłącznie według kryteriów technicznych ustalanych przez dostawcę oprogramowania.

Sprawiedliwość algorytmiczna wymaga legitymacji społecznej i proceduralnej

Ekonomia instytucjonalna pozwala analizować, w jaki sposób konkretne wskaźniki efektywności zostają utrwalone przez system bodźców organizacyjnych. Jeśli instytucja nagradza przede wszystkim szybkość rozpatrywania spraw, może uznać technologię zwiększającą przepustowość procedur za skuteczną, nawet jeśli pogarsza ona jakość indywidualnej oceny. Gdy przedsiębiorstwo premiuje redukcję kosztów obsługi, może ograniczać zatrudnienie specjalistów niezbędnych do wykrywania błędów algorytmicznych. Z kolei platforma maksymalizująca zaangażowanie użytkowników może preferować treści wywołujące silne reakcje emocjonalne, ignorując ich wartość poznawczą oraz społeczne koszty polaryzacji.

Problem ten wiąże się z teorią pryncypała i agenta oraz analizą asymetrii informacji. Projektant systemu, dostawca technologii, instytucja zamawiająca i użytkownik końcowy nie muszą mieć identycznych celów ani równorzędnych możliwości oceny rezultatów. Dlatego sama deklaracja bezstronności algorytmu nie może zastępować niezależnej kontroli jego rzeczywistych skutków. Nie oznacza to jednak, że wszystkie decyzje algorytmiczne powinny zostać bezrefleksyjnie zastąpione oceną ludzką. Badania nad oceną ekspercką pokazują, że ludzie również popełniają systematyczne błędy, ulegają stereotypom, błędnie interpretują prawdopodobieństwa i pod wpływem okoliczności niezwiązanych z meritum sprawy podejmują niewłaściwe decyzje. W niektórych zadaniach dobrze zaprojektowany model statystyczny może przewyższać przeciętną skuteczność człowieka, zwiększać spójność kryteriów i ograniczać arbitralność. Wartość takiego rozwiązania zależy jednak od tego, czy porównanie opiera się na realistycznych alternatywach, a nie na idealizowanym obrazie nieomylnego eksperta lub doskonałego algorytmu. Należy zatem badać rzeczywiste koszty błędów, możliwość uczenia się na doświadczeniach, procedury odwoławcze oraz warunki funkcjonowania modelu wewnątrz instytucji.

Odpowiedzialna automatyzacja nie polega na zastąpieniu ludzkiego osądu matematycznymi wskaźnikami, lecz na takim zorganizowaniu relacji między wiedzą ekspercką a predykcją, aby ograniczać słabości obu tych metod. W perspektywie politycznej kluczowy pozostaje problem demokratycznej legitymacji kryteriów sprawiedliwości wykorzystywanych przez systemy AI. Wybór między parytetem demograficznym, równością szans, kalibracją ryzyka a minimalizacją różnych rodzajów błędów wpływa na sytuację całych populacji. Nie jest to zatem wyłącznie zagadnienie techniczne, które mogłoby zostać rozstrzygnięte przez wąską grupę specjalistów.

Odpowiedzialność demokratyczna wymaga określenia, kto jest uprawniony do definiowania celu, jakie wartości powinny być chronione, które kompromisy są dopuszczalne i w jaki sposób osoby

dotknięte decyzją mogą kwestionować jej podstawy. W pewnych obszarach rozstrzygnięcia te wynikają z norm prawnych, w innych zaś wymagają debaty publicznej i procedur regulacyjnych. W każdym przypadku konieczne jest odróżnienie matematycznej możliwości realizacji danego kryterium od społecznej legitymacji jego przyjęcia. Istotne jest również rozpoznanie, że formalna równość modelu nie gwarantuje rzeczywistej równości uczestników procesu. Dwie osoby z identycznym wynikiem predykcyjnym mogą mieć zupełnie inne możliwości zakwestionowania decyzji: jedna skorzysta z pomocy prawnika lub eksperta, druga pozostanie bez środków pozwalających zrozumieć podstawy rozstrzygnięcia.

Z perspektywy teorii sprawiedliwości proceduralnej sama jednolitość technicznych reguł nie wyczerpuje wymogów sprawiedliwego traktowania. Konieczne jest uwzględnienie dostępności informacji, prawa do rzetelnego uzasadnienia, realnych środków odwoławczych i możliwości naprawienia szkody. To właśnie tutaj matematyczna analiza predykcji spotyka się z socjologicznym badaniem nierówności zasobów i prawną ochroną podmiotowości jednostki. Problematyka ta pozwala rozwinąć teorię epistemicznej niesprawiedliwości Mirandy Fricker, dotyczącą sytuacji, w których człowiek doświadcza krzywdy jako podmiot wiedzy. Niesprawiedliwość świadectwa pojawia się wtedy, gdy osobie przypisuje się nieuzasadnienie mniejszą wiarygodność ze względu na stereotypy społeczne; niesprawiedliwość hermeneutyczna dotyczy natomiast nierównego dostępu do zasobów pojęciowych umożliwiających interpretację własnego doświadczenia. System algorytmiczny może uczestniczyć w tych procesach, jeśli automatycznie deprecjonuje określone formy wypowiedzi, doświadczenia lub wiedzę lokalną. Może on również pogłębiać asymetrię poznawczą między instytucją a obywatelem, który nie dysponuje językiem ani informacjami pozwalającymi opisać przyczynę niekorzystnej decyzji. Nie wymaga to przypisywania maszynie świadomego zamiaru dyskryminacji.

Wystarczy, że jej wyniki zostaną włączone w relacje społeczne, w których określone świadectwa i doświadczenia są systematycznie marginalizowane. Z tej perspektywy stronniczość algorytmiczna nie jest jedynie defektem technologii, lecz szczególnym sposobem utrwalania relacji władzy przez infrastrukturę poznawczą. Instytucja dysponująca modelem decyzyjnym może zyskać przewagę informacyjną nad osobami ocenianymi, a matematyczna złożoność procedury może utrudniać krytyczną kontrolę rezultatów.

Nie oznacza to, że każda asymetria wiedzy jest nieuzasadniona – specjalizacja naukowa i techniczna jest cechą nowoczesnego społeczeństwa. Problem pojawia się jednak wtedy, gdy przewaga specjalistyczna staje się substytutem odpowiedzialności, a obywatel traci praktyczną możliwość zakwestionowania decyzji wpływającej na jego prawa i interesy.

Sprawiedliwość AI wymaga uznania granic matematycznej reprezentacji człowieka

Odpowiedzialne systemy AI powinny nie tylko zwiększać skuteczność instytucji, ale przede wszystkim umożliwiać kontrolę nad tym, w jaki sposób ta skuteczność została osiągnięta. W tym miejscu ujawnia się zasadnicze znaczenie metafory wychowania sztucznej inteligencji. Jeśli rozumieć ją jako proces stopniowego kształtowania zdolności systemu do działania zgodnie z uzasadnionymi celami człowieka, to matematyczna poprawność predykcji jest zaledwie jednym z warunków dojrzałości technologicznej. Równie istotne okazują się umiejętność rozpoznawania własnych ograniczeń, właściwe reagowanie na niepewność, unikanie nieproporcjonalnych szkód oraz podatność na ludzką korektę. Nie należy jednak przypisywać takim systemom moralnej dojrzałości w znaczeniu właściwym osobie ludzkiej. Techniczne dostrajanie modelu nie jest tożsame z wewnętrznym rozwojem sumienia, odpowiedzialności czy empatii. Istota proponowanej przez autora analogii polega na podkreśleniu obowiązków konstruktorów i opiekunów instytucjonalnych, którzy powinni tworzyć warunki bezpiecznego rozwoju kompetencji systemu, nie rezygnując przy tym z własnej odpowiedzialności za jego wykorzystanie.

Analiza ta obnaża sprzeczność współczesnej cywilizacji cyfrowej. Społeczeństwa coraz skuteczniej przekładają rozległe fragmenty rzeczywistości na reprezentacje matematyczne, lecz sama możliwość obliczeniowego modelowania nie dowodzi, że rozpoznano wszystkie właściwości istotne dla ludzkiego życia. Człowiek może zostać sklasyfikowany jako jednostka o określonym poziomie ryzyka, pracownik o przewidywanej produktywności, pacjent o oszacowanym koszcie leczenia albo konsument o wyznaczonej skłonności do zakupu.

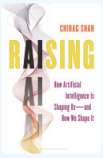
Żadna z tych reprezentacji nie stanowi jednak pełnego opisu jego podmiotowości, możliwości rozwojowych, historii społecznej ani normatywnego znaczenia jego praw. Modele mogą dostarczać cennych informacji, ale ich użyteczność nie uzasadnia utożsamiania człowieka z rezultatem klasyfikacji. Rozpoznanie granic matematycznej reprezentacji staje się tym samym jednym z najważniejszych warunków odpowiedzialnego rozwoju sztucznej inteligencji. Nauka nie powinna rezygnować z dążenia do coraz dokładniejszych modeli, gdyż ich osiągnięcia mogą zwiększać bezpieczeństwo, poprawiać diagnostykę, ograniczać arbitralność i wspierać dostęp do wiedzy. Powinna jednak konsekwentnie oddzielać skuteczność przewidywania od prawdziwości wyjaśnienia przyczynowego oraz od moralnej zasadności wykorzystania predykcji. Zastosowanie AI wobec człowieka wymaga bowiem nie tylko odpowiedzi na pytanie o to, co prawdopodobnie się wydarzy, lecz także o to, dlaczego określone zdarzenie może nastąpić, jakie działania pozostają dopuszczalne i jakie konsekwencje powinny być chronione przed automatycznym urzeczywistnieniem.

Matematyka uprzedzeń ujawnia więc ograniczenia zarówno naiwnego technokratyzmu, jak i odruchowego odrzucania algorytmicznych metod podejmowania decyzji. Pierwszy zakłada, że doskonałość obliczeniowa nieuchronnie prowadzi do sprawiedliwości; drugi ignoruje możliwość wykorzystywania naukowych modeli do identyfikowania i ograniczania nierówności. Bardziej wymagające stanowisko opiera się na uznaniu, że sprawiedliwość algorytmiczna pozostaje problemem jednocześnie statystycznym, przyczynowym, instytucjonalnym i aksjologicznym. Jej badanie wymaga nie tylko doskonalenia modeli, lecz także poprawnego definiowania celów, rozpoznawania genezy danych, analizowania skutków decyzji, zapewniania procedur odwoławczych i prowadzenia publicznej refleksji nad dopuszczalnymi granicami automatyzacji. W tej perspektywie cicha produkcja nierówności okazuje się szczególnie niebezpieczna właśnie dlatego, że nie musi naruszać wewnętrznej logiki systemu technicznego. Algorytm może działać zgodnie z projektem, organizacja realizować przyjęte wskaźniki efektywności, a system kontroli jakości potwierdzać poprawność obliczeń, podczas gdy społeczne konsekwencje pozostają niezbadane albo zostają uznane za uboczne.

Zrozumienie tej zależności wymaga przesunięcia uwagi z samego modelu ku całemu procesowi, w którym predykcja uzyskuje status decyzji dotyczącej konkretnej osoby. Dopiero takie spojrzenie pozwala przejść od matematycznej analizy nierówności wyników do zbadania rzeczywistych szkód społecznych, odpowiedzialności instytucjonalnej i granic prawnej ochrony człowieka w środowisku coraz bardziej zautomatyzowanym.

> Czy w świecie, w którym człowiek zostaje sprowadzony do wyniku klasyfikacji ryzyka lub produktywności, matematyczna precyzja nie staje się nową formą niewidzialnego zniewolenia? Prawdziwym wyzwaniem nie jest stworzenie modelu idealnie sprawiedliwego, lecz uznanie momentu, w którym algorytm musi ustąpić miejsca ludzkiej podmiotowości. Ostatecznie to nie kod, lecz nasza odwaga do zakwestionowania „obiektywnych” danych zadecyduje o tym, czy AI będzie narzędziem emancypacji, czy jedynie wydajniejszym mechanizmem reprodukcji niesprawiedliwości.

Inspiracje

[1] 

"Raising AI" Chirag Shah

Bloomsbury Academic | ISBN: 9798216452409

Chirag Shah jest profesorem w Szkole Informatyki i profesorem nadzwyczajnym w Szkole Informatyki i Inżynierii im. Paula G. Allena na Uniwersytecie Waszyngtońskim. Uzyskał doktorat z informatyki na Uniwersytecie Karoliny Północnej w Chapel Hill, a także dyplomy z IIT Madras i UMass Amherst. Shah jest dyrektorem założycielem InfoSeeking Lab i współdyrektorem założycielem Center for Responsibility in AI Systems & Experiences (RAISE). Jego badania koncentrują się na inteligentnych systemach dostępu do informacji, w tym wyszukiwaniu, rekomendacjach, systemach konwersacyjnych i agentowej sztucznej inteligencji, kładąc nacisk na uczciwość, przejrzystość i odpowiedzialną sztuczną inteligencję. Jako wybitny członek ACM i ASIS&T, pełnił funkcję konsultanta Organizacji Narodów Zjednoczonych i intensywnie współpracował z branżowymi organizacjami badawczymi w zakresie etycznych i zorientowanych na użytkownika architektur uczenia maszynowego.

Chirag Shah *is a Professor at the Information School and an Adjunct Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. He earned his Ph.D. in Information Science from the University of North Carolina at Chapel Hill, along with degrees from IIT Madras and UMass Amherst. Shah is the Founding Director of the InfoSeeking Lab and Founding Co-Director of the Center for Responsibility in AI Systems & Experiences (RAISE). His research focuses on intelligent information access systems, including search, recommendation, conversational systems, and agentic AI, emphasizing fairness, transparency, and responsible AI. A Distinguished Member of ACM and ASIS&T, he has served as a consultant for the United Nations and worked extensively with industry research organizations on ethical and user-centered machine learning architectures.*

University of Washington

Literatura uzupełniająca

Książki

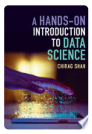
Książki autora i dzieła przywołane w tekście:



[1] "Agent Nation"

Chirag Shah

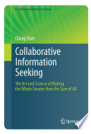
2026 | Springer Nature | ISBN: 9798868824548



[2] "A Hands-On Introduction to Data Science"

Chirag Shah

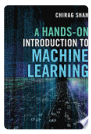
2020 | Cambridge University Press | ISBN: 9781108472449



[3] "Collaborative Information Seeking"

Chirag Shah

2012 | Springer Science & Business Media | ISBN: 9783642288135



[4] "A Hands-On Introduction to Machine Learning"

Chirag Shah

2022 | Cambridge University Press | ISBN: 9781009123303



[5] "Task Intelligence for Search and Recommendation"

Chirag Shah, Ryen W. White

2022 | Springer Nature | ISBN: 9783031023262



[6] "Interactive IR User Study Design, Evaluation, and Reporting"

Jiqun Liu, Chirag Shah

2022 | Springer Nature | ISBN: 9783031023194

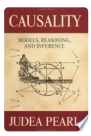
Książki autorów przywoływanych i literatura pokrewna:



[7] "Mathematical Statistics"

Peter J. Bickel, Kjell A. Doksum

1977



[8] "Causality: Models, Reasoning and Inference"

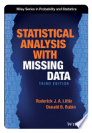
Judea Pearl

Cambridge University Press | ISBN: 9780521773621

[9] "Eight Myths About Causality and Structural Models"

Kenneth Bollen, Judea Pearl

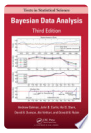
2013



[10] "Statistical analysis with missing data"

Donald Rubin

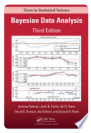
2019 | John Wiley & Sons | ISBN: 9780470526798



[11] "Bayesian Data Analysis"

Donald Rubin

CRC Press | ISBN: 9781439898208



[12] "Bayesian Data Analysis, Third Edition"

Donald Rubin

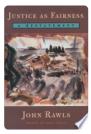
CRC Press | ISBN: 9781040062340



[13] "Scarcity: Why Having Too Little Means So Much"

Sendhil Mullainathan

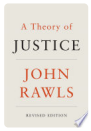
Penguin UK | ISBN: 9780141961194



[14] "Justice as Fairness"

John Rawls

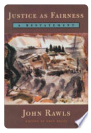
Belknap Press | ISBN: 9780674005112



[15] "A Theory of Justice"

John Rawls

2020 | Harvard University Press | ISBN: 9780674257672



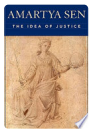
[16] "Justice as Fairness: A Restatement"

John Rawls

Harvard University Press | ISBN: 9780674005105

[17] "India: Economic Development and Social Opportunity"

Amartya Sen



[18] "The Idea of Justice"

Amartya Sen

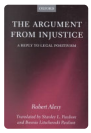
Harvard University Press | ISBN: 9780674036130



[19] "The Alexy-Poscher Debate on Legal Principles"

Robert Alexy, Ralf Poscher

2025 | Bloomsbury Publishing | ISBN: 9781509980604



[20] "The Argument from Injustice"

Robert Alexy

2002 | OUP Oxford | ISBN: 9780198259879



[21] "A Theory of Legal Argumentation"

Robert Alexy

1989



[22] "Epistemic Injustice: Power and the Ethics of Knowing"

Miranda Fricker

Oxford University Press on Demand | ISBN: 9780199570522



[23] "Epistemic Injustice"

Miranda Fricker

2007 | Oxford University Press | ISBN: 9780198237907

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Individu et justice sociale"

John Rawls

1988 | Individu et justice sociale | DOI: 10.3917/ls.terre.1988.01.0277

Źródło

[2] "Adversarial examples from computational constraints"

Sébastien Bubeck, Eric Price, Ilya Razenshteyn

2018 | arXiv | DOI: 10.48550/arxiv.1805.10204

Źródło

[3] "Approximate compression: enhancing compressibility through data approximation"

Harini Suresh, Shashank Hegde, John Sartori

2017 | DOI: 10.1145/3139315.3139319

Źródło

[4] "Novel Intra-Prediction Framework for H.264 Video Compression using Decision and Prediction Mode"

Pradeep Kumar, Harini Suresh

2017 | International Journal of Advanced Computer Science and Applications | DOI: 10.14569/ijacsa.2017.080409

Źródło

[5] "Multi-Mode Schema for an Efficient Intra-Frame Prediction Technique for H.264"

Pradeep Kumar N.S, Harini Suresh

2015 | HAL

Źródło

[6] "Encoding time optimization for intra-frame reconstruction schemes for H.264"

Pradeep Kumar N.S, Harini Suresh

2018 | International Journal of Engineering & Technology | DOI: 10.14419/ijet.v7i2.33.14161

Źródło

[7] "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle"

Harini Suresh, John V. Guttag

2021 | ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO) | DOI: 10.1145/3465416.3483305

Źródło

[8] "An Unsupervised Learning Model for Deformable Medical Image Registration"

Guha Balakrishnan, Amy Zhao, Mert Rory Sabuncu, Adrian Vasile Dalca, John V. Guttag

2018 | IEEE Computer Society Conference on Computer Vision and Pattern Recognition/Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition/Proceedings | DOI: 10.1109/cvpr.2018.00964

Źródło

[9] "Detecting Pulse from Head Motions in Video"

Guha Balakrishnan, Frédo Durand, John V. Guttag

2013 | IEEE Computer Society Conference on Computer Vision and Pattern Recognition/Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition/Proceedings | DOI: 10.1109/cvpr.2013.440

Źródło

[10] "Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle"

Harini Suresh, John Guttag

2021 | MIT Case Studies in Social and Ethical Responsibilities of Computing | DOI: 10.21428/2c646de5.c16a07bb

Źródło

[11] "Measurement of Diversity"

Edward Simpson

1949 | Nature | DOI: 10.1038/163688a0

Źródło

[12] "Edward Simpson: Bayes at Bletchley Park"

Edward Simpson

2010 | Significance | DOI: 10.1111/j.1740-9713.2010.00424.x

Źródło

[13] "DISCUSSION: Statistical Investigation of Regional Structure and Reactions"

Paul B. Simpson

1956 | Papers in Regional Science | DOI: 10.1111/j.1435-5597.1956.tb01562.x

Źródło

[14] "Simultaneous analysis of Lasso and Dantzig selector"

Peter J. Bickel, Ya'acov Ritov, Alexandre B. Tsybakov

2009 | The Annals of Statistics | DOI: 10.1214/08-aos620

Źródło

[15] "A nonparametric view of network models and Newman–Girvan and other modularities"

Peter J. Bickel, Aiyou Chen

2009 | Proceedings of the National Academy of Sciences | DOI: 10.1073/pnas.0907096106

Źródło

[16] "Obstacles to High-Dimensional Particle Filtering"

Chris Snyder, Thomas Bengtsson, Peter J. Bickel, J. G. Anderson

2008 | Monthly Weather Review | DOI: 10.1175/2008mwr2529.1

Źródło

[17] **"On Some Global Measures of the Deviations of Density Function Estimates"**

Peter J. Bickel, M. Rosenblatt

1973 | The Annals of Statistics | DOI: 10.1214/aos/1176342558

Źródło

[18] **"Sex Bias in Graduate Admissions: Data from Berkeley"**

Peter J. Bickel, Eugene A. Hammel, J. W. O'Connell

1975 | Science | DOI: 10.1126/science.187.4175.398

Źródło

[19] **"Comparing Household Structure Over Time and Between Cultures"**

Eugene A. Hammel, Peter Laslett

1974 | Comparative Studies in Society and History | DOI: 10.1017/s0010417500007362

Źródło

[20] **"Fertility Decline in Prussia, 1875–1910: A Pooled Cross-Section Time Series Analysis"**

Patrick R. Galloway, Eugene A. Hammel, Ronald Demos Lee

1994 | Population Studies | DOI: 10.1080/0032472031000147516

Źródło

[21] **"Alternative social structures and ritual relations in the Balkans"**

Eugene A. Hammel

1968 | Prentice Hall eBooks

Źródło

[22] **"Application of the SSMRP methodology to the seismic risk at the Zion Nuclear Power Plant"**

M.P. Bohn, L.C. Shieh, J.E. Wells, L.E. Cover, D.L. Bernreuter, W.J. O'Connell

1983 | OSTI OAI (U.S. Department of Energy Office of Scientific and Technical Information) | DOI: 10.2172/5583144

Źródło

[23] **"2H(d, γ)4He reaction from 6 to 19 MeV"**

Walter E. Meyerhof, W. C. Feldman, Stanley Gilbert, W.J. O'Connell

1969 | Nuclear Physics A | DOI: 10.1016/0375-9474(69)90590-9

Źródło

[24] **"Spectrum of γ rays in the 8- to 20-MeV range from Cf252 spontaneous fission"**

Franz Dietrich, John C. Browne, W.J. O'Connell, M. J. Kay

1974 | Physical Review C | DOI: 10.1103/physrevc.10.795

Źródło

[25] **"Waste package performance assessment: Deterministic system model, program scope and specification"**

CA (United States) Lawrence Livermore National Lab., W.J. O'Connell, Robert S. Drach

1986 | OSTI OAI (U.S. Department of Energy Office of Scientific and Technical Information) | DOI: 10.2172/60117

Źródło

[26] **"Bayesian Networks : a Model of Self-Activated Memory for Evidential Reasoning"**

Judea Pearl

1985 | Conference Cognitive Science

Źródło

[27] **"From Adams' conditionals to default expressions, causal conditionals, and counterfactuals"**

Judea Pearl

1994 | Cambridge University Press eBooks

Źródło

[28] **"Testing regression models with fewer regressors"**

Judea Pearl, Peyman Meshkat

1999 | eScholarship (California Digital Library)

Źródło

[29] **"Causal Inference under Interference and Model Uncertainty"**

Chi Zhang, Karthika Mohan, Judea Pearl

2023 | Proceedings of Machine Learning Research

Źródło

[30] **"Replication data for: Bayesian Data Analysis, Second Edition"**

Andrew Gelman, John B. Carlin, Hal S. Stern, And Donald B. Rubin

2009 | Harvard Dataverse | DOI: 10.7910/dvn/htb4hj

Źródło

[31] "Rubin Causal Model"

Donald B. Rubin

2025 | International Encyclopedia of Statistical Science | DOI: 10.1007/978-3-662-69359-9_540

Źródło

[32] "Statistical Inference for Causal Effects, with Emphasis on Applications in Psychometrics and Education"

Donald B. Rubin

The SAGE Handbook of Innovation in Social Research Methods | DOI: 10.4135/9781446268261.n29

Źródło

[33] "Comment: Self-Experimentation for Causal Effects"

Donald B. Rubin

2001 | CHANCE | DOI: 10.1080/09332480.2001.10542261

Źródło

[34] "Fairness through awareness"

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard S. Zemel

2012 | Conference on Innovations in Theoretical Computer Science | DOI: 10.1145/2090236.2090255

Źródło

[35] "Understanding deep learning requires rethinking generalization"

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals

2016 | arXiv | DOI: 10.48550/arxiv.1611.03530

Źródło

[36] "Understanding deep learning (still) requires rethinking generalization"

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals

2021 | Communications of the ACM | DOI: 10.1145/3446776

Źródło

[37] "Test-Time Training with Self-Supervision for Generalization under Distribution Shifts"

Yu Sun, Xiaolong Wang, Zhuang Liu, John J. Miller, Alexei A. Efros, Moritz Hardt

2020 | International Conference on Machine Learning | DOI: 10.48550/arxiv.1909.13231

Źródło

[38] "Phi-4 Technical Report"

Marah Abdin, Jyoti Aneja, Harkirat Singh Behl, Sébastien Bubeck, Ronen Eldan, Eric Price

2024 | arXiv | DOI: 10.48550/arxiv.2412.08905

Źródło

[39] "Compressed Sensing using Generative Models"

Ashish Bora, Ajil Jalal, Eric Price, Alexandros G. Dimakis

2017 | arXiv | DOI: 10.48550/arxiv.1703.03208

Źródło

[40] "AmbientGAN: Generative models from lossy measurements"

Ashish Bora, Eric Price, Alexandros G. Dimakis

2018 | International Conference on Learning Representations

Źródło

[41] "Recursive Models for Long-Horizon Reasoning"

Chen-Xiao Yang, N. Srebro, Zhi-Yuan Li

2026 | arXiv.org | DOI: 10.48550/arxiv.2603.02112

Źródło

[42] "Invited Open Problem: Is the Power of Deep Learning over Linear Models Inherently Distribution Dependent?"

V. Feldman, Pritish Kamath, N. Srebro

2026 | Annual Conference Computational Learning Theory

Źródło

[43] "Recursive Language Models Generalize Out of Domain"

Chen-Xiao Yang, Zhi-Yuan Li, D. McAllester, N. Srebro

2026 | arXiv.org | DOI: 10.48550/arxiv.2609.20831

Źródło

[44] "Maximizing the spread of influence through a social network"

David Kempe, Jon Michael Kleinberg, Eva Tardos

2003 | ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) | DOI: 10.1145/956750.956769

Źródło

[45] "Authoritative sources in a hyperlinked environment"

Jon Michael Kleinberg

1999 | Journal of the ACM | DOI: 10.1145/324133.324140

Źródło

[46] "The link-prediction problem for social networks"

David Liben-Nowell, Jon Michael Kleinberg

2007 | Journal of the American Society for Information Science and Technology | DOI: 10.1002/asi.20591

Źródło

[47] "Graphs over time: densification laws, shrinking diameters and possible explanations"

Jure Leskovec, Jon Michael Kleinberg, Christos Faloutsos

2005 | ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) | DOI: 10.1145/1081870.1081893

Źródło

[48] "Dissecting racial bias in an algorithm used to manage the health of populations"

Ziad Obermeyer, Brian W. Powers, Christine Vogeli, Sendhil Mullainathan

2019 | Science | DOI: 10.1126/science.aax2342

Źródło

[49] "Inherent Trade-Offs in the Fair Determination of Risk Scores"

Jon Michael Kleinberg, Sendhil Mullainathan, Manish Raghavan

2016 | DROPS (Schloss Dagstuhl – Leibniz Center for Informatics) | DOI: 10.4230/lipics.itcs.2017.43

Źródło

[50] "Do People Mean What They Say? Implications for Subjective Survey Data"

Marianne Bertrand, Sendhil Mullainathan

2001 | American Economic Review | DOI: 10.1257/aer.91.2.67

Źródło

[51] "Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability"

Jon Kleinberg, Sendhil Mullainathan

2019 | DOI: 10.3386/w25854

Źródło

[52] "Mitigating bias in algorithmic hiring: evaluating claims and practices"

Manish Raghavan, Solon Barocas, Jon Michael Kleinberg, Karen E. C. Levy

2020 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3351095.3372828

Źródło

[53] "On Fairness and Calibration"

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Michael Kleinberg, Kilian Q. Weinberger

2017 | neural information processing systems | DOI: 10.48550/arxiv.1709.02012

Źródło

[54] "Roles for computing in social change"

Rediet Abebe, Solon Barocas, Jon Michael Kleinberg, Karen E. C. Levy, Manish Raghavan

2020 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3351095.3372871

Źródło

[55] "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments"

Alexandra Chouldechova

2017 | Big Data | DOI: 10.1089/big.2016.0047

Źródło

[56] "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting"

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova

2019 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3287560.3287572

Źródło

[57] "The Frontiers of Fairness in Machine Learning"

Alexandra Chouldechova, Aaron Roth

2018 | arXiv | DOI: 10.48550/arxiv.1810.08810

Źródło

[58] "A snapshot of the frontiers of fairness in machine learning"

Alexandra Chouldechova, Aaron Roth

2020 | Communications of the ACM | DOI: 10.1145/3376898

Źródło

[59] "Justice as Fairness: Political Not Metaphysical"

John Rawls

1991 | Equality and Liberty | DOI: 10.1007/978-1-349-21763-2_10

Źródło

[60] "Analysis of Pyrimidine Catabolism in Drosophila melanogaster Using Epistatic Interactions With Mutations of Pyrimidine Biosynthesis and β-Alanine Metabolism"

J. Rawls

2006 | Genetics | DOI: 10.1534/genetics.105.052753

Źródło

[61] "What do We Want From a Theory of Justice? *"

Sen Amartya

2017 | Theories of Justice | DOI: 10.4324/9781315236322-4

Źródło

[62] "Evidence and Efficacy in the Era of Digital Care"

S. Gondi, Brian W. Powers, W. Shrank

2022 | Journal of general internal medicine | DOI: 10.1007/s11606-022-07445-0

Źródło

[63] "Quasi-Experimentation: Design & Analysis Issues for Field Settings"

Thomas D. Cook, Donald T. Campbell

1979

Źródło

[64] "Common fate, similarity, and other indices of the status of aggregates of persons as social entities"

Donald T. Campbell

2007 | Systems Research and Behavioral Science | DOI: 10.1002/bs.3830030103

Źródło

[65] "Regression-discontinuity analysis: An alternative to the ex post facto experiment."

Donald L. Thistlethwaite, Donald T. Campbell

1960 | Journal of Educational Psychology | DOI: 10.1037/h0044319

Źródło

[66] "Factors relevant to the validity of experiments in social settings."

Donald T. Campbell

1957 | Psychological Bulletin | DOI: 10.1037/h0040950

Źródło

[67] "Runaway Feedback Loops in Predictive Policing"

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, Suresh Venkatasubramanian

2018

Źródło

[68] "Decision making with limited feedback"

Danielle Ensign, Frielder Sorelle, Neville Scott, Carlos Scheidegger, Suresh Venkatasubramanian

2018 | Algorithmic Learning Theory

Źródło

[69] "Decision making with limited feedback : Error bounds for recidivism prediction and predictive policing"

Danielle Ensign, Scott Neville

2017

Źródło

[70] "On the Phase Diagram of the Polymer Model"

A. Krishnan, S. Mkrtchyan, Scott Neville

2024 | Journal of theoretical probability | DOI: 10.1007/s10959-026-01504-6

Źródło

[71] "A comparative study of fairness-enhancing interventions in machine learning"

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton

2019 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3287560.3287589

Źródło

[72] "Problems with Shapley-value-based explanations as feature importance measures"

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, Sorelle A. Friedler

2020 | arXiv | DOI: 10.48550/arxiv.2002.11097

Źródło

[73] "VisTrails: visualization meets data management"

Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos Scheidegger, Claudio T. Silva

2006 | ACM SIGMOD International Conference on Management of Data (SIGMOD) | DOI: 10.1145/1142473.1142574

Źródło

[74] "On the (im)possibility of fairness"

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian

2016 | arXiv | DOI: 10.48550/arxiv.1609.07236

Źródło

[75] "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity"

Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian

2007 | IEEE International Conference on Data Engineering (ICDE) | DOI: 10.1109/icde.2007.367856

Źródło

[76] "Fairness and Abstraction in Sociotechnical Systems"

Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, Janet Vertesi

2019 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3287560.3287598

Źródło

[77] "Differential Privacy Has Disparate Impact on Model Accuracy"

Eugene Bagdasaryan, Vitaly Shmatikov

2019 | neural information processing systems | DOI: 10.48550/arxiv.1905.12101

Źródło

[78] "Blind Backdoors in Deep Learning Models"

Eugene Bagdasaryan, Vitaly Shmatikov

2020 | arXiv | DOI: 10.48550/arxiv.2005.03823

Źródło

[79] "Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures"

Eugene Bagdasaryan, Vitaly Shmatikov

2022 | IEEE Symposium on Security and Privacy (S&P) | DOI: 10.1109/sp46214.2022.9833572

Źródło

[80] **"Operationalizing Contextual Integrity in Privacy-Conscious Assistants"**

Sahra Ghalebikesabi, Eugene Bagdasaryan, Ren Yi, Itay Yona, Ilia Shumailov

2024 | arXiv | DOI: 10.48550/arxiv.2408.02373

Źródło

[81] **"Self-Supervised Learning of Point Clouds via Orientation Estimation"**

Omid Poursaeed, Tianxing Jiang, Han Qiao, Nayun Xu, Vladimir G. Kim

2020 | International Conference on 3D Vision (3DV) | DOI: 10.1109/3dv50981.2020.00112

Źródło

[82] **"Coupling Explicit and Implicit Surface Representations for Generative 3D Modeling"**

Omid Poursaeed, Matthew Fisher, Noam Aigerman, Vladimir G. Kim

2020 | Lecture notes in computer science | DOI: 10.1007/978-3-030-58607-2_39

Źródło

[83] **"A run shock-erosion model"**

Mohammad Hossein Poursaeed

2019 | Communications in Statistics - Theory and Methods | DOI: 10.1080/03610926.2019.1649425

Źródło

[84] **"On δ-shock model in a multi-state system"**

Mohammad Hossein Poursaeed

2019 | Communications in Statistics - Theory and Methods | DOI: 10.1080/03610926.2019.1565784

Źródło

[85] **"Membership Inference Attacks Against Machine Learning Models"**

Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov

2017 | IEEE Symposium on Security and Privacy (S&P) | DOI: 10.1109/sp.2017.41

Źródło

[86] **"Privacy-Preserving Deep Learning"**

Reza Shokri, Vitaly Shmatikov

2015 | ACM SIGSAC Conference on Computer and Communications Security (CCS) | DOI: 10.1145/2810103.2813687

Źródło

[87] "De-anonymizing Social Networks"

Arvind Narayanan, Vitaly Shmatikov

2009 | IEEE Symposium on Security and Privacy (S&P) | DOI: 10.1109/sp.2009.22

Źródło

[88] "Machine Learning Models that Remember Too Much"

Congzheng Song, Thomas Ristenpart, Vitaly Shmatikov

2017 | ACM SIGSAC Conference on Computer and Communications Security (CCS) | DOI: 10.1145/3133956.3134077

Źródło

[89] "El Imperio de la justicia: de la teoría general del derecho, de las decisiones e interpretaciones de los jueces y de la integridad política y legal como clave de la teoría y la práctica"

R. Dworkin

1988 | Dialnet (Universidad de la Rioja)

Źródło

[90] "The right to Sabbath: an interpretation of the legal practice of the Sabbath from the alternative theory of Ronald Dworkin"

R. Dworkin, Cristian David Ocampo-Macías

2019 | DOI: 10.18041/1794-7200/criteriojuridico.2

Źródło

[91] "On the Structure of Legal Principles"

Robert Alexy

2000 | Ratio Juris | DOI: 10.1111/1467-9337.00157

Źródło

[92] "Balancing, constitutional review, and representation"

Robert Alexy

2005 | International Journal of Constitutional Law | DOI: 10.1093/icon/moi040

Źródło

[93] "A Theory of Legal Argumentation: The Theory of Rational Discourse as Theory of Legal Justification"

Robert Alexy, Ruth Adler, Neil MacCormick

1989

Źródło

[94] **"Justification and Application of Norms*"**

Robert Alexy

1993 | Ratio Juris | DOI: 10.1111/j.1467-9337.1993.tb00144.x

Źródło

[95] **"Epistemic Injustice: Power and the Ethics of Knowing"**

Miranda Fricker

2007 | DOI: 10.1093/acprof:oso/9780198237907.001.0001

Źródło

[96] **"Evolving Concepts of Epistemic Injustice"**

Miranda Fricker

2017 | Routledge eBooks | DOI: 10.4324/9781315212043-5

Źródło

[97] **"Epistemic justice as a condition of political freedom?"**

Miranda Fricker

2012 | Synthese | DOI: 10.1007/s11229-012-0227-3

Źródło

[98] **"IX—Rational Authority and Social Power: Towards a Truly Social Epistemology"**

Miranda Fricker

1998 | Proceedings of the Aristotelian Society (Hardback) | DOI: 10.1111/1467-9264.00030

Źródło

Artykuły naukowe autora:

[99] **"Raising AI: How Artificial Intelligence Is Shaping Us—and How We Shape It"**

Chirag Shah

2026 | DOI: 10.5040/9798216452430

Źródło

[100] **"Ship It, Then Apologize: We Can Do Better Than This for AI Advancements"**

Chirag Shah

2026 | SSRN Electronic Journal | DOI: 10.2139/ssrn.6952240

Źródło

Treść tworzy Fundacja Dobre Państwo.
Redaguje i publikuje APA ONE, autorski system redakcyjny Fundacji oparty na AI.
APA ONE Publishing System
Fundacja Dobre Państwo © 2026
Article ID: e759f06d-0e3c-4f45-ad70-caa6c6e85793