Wychowanie sztucznej inteligencji: etyka odpowiedzialności w obliczu cichej stronniczości algorytmów na podstawie książki Chiraga Shaha Raising AI

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje problem asymetrii między gwałtownym rozwojem mocy obliczeniowej AI a brakiem odpowiednich mechanizmów nadzoru społecznego i etycznego. Opierając się na książce Chiraga Shaha „Raising AI”, autor wskazuje, że największym zagrożeniem nie jest futurystyczny bunt maszyn, lecz „cicha stronniczość” – rutynowe przekształcanie niedoskonałych danych w decyzje instytucjonalne. Systemowa niesprawiedliwość często wynika nie ze złej woli twórców, lecz z błędnych funkcji celu i braku reprezentatywności danych. Tekst podkreśla konieczność przejścia od lęków dystopijnych ku konkretnej odpowiedzialności projektantów i instytucji, traktując rozwój AI jako proces „wychowania” systemu w środowisku społecznym, gdzie efektywność statystyczna nie może zastąpić sprawiedliwości i moralnego osądu.

The article analyzes the problem of asymmetry between the rapid development of AI computing power and the lack of appropriate social and ethical oversight mechanisms. Drawing on Chirag Shah's book 'Raising AI,' the author indicates that the greatest threat is not a futuristic machine rebellion, but 'silent bias'—the routine transformation of imperfect data into institutional decisions. Systemic injustice often results not from the ill will of creators, but from flawed objective functions and lack of data representativeness. The text emphasizes the need to move from dystopian fears toward concrete responsibility for designers and institutions, treating AI development as a process of 'raising' the system within a social environment, where statistical efficiency cannot replace justice and moral judgment.

Artykuł analizuje krytyczny rozdźwięk między gwałtownym wzrostem mocy obliczeniowej sztucznej inteligencji a brakiem dojrzałości instytucjonalnej w nadzorowaniu jej skutków społecznych. Autor stawia tezę, że współczesna lękowa fascynacja futurystyczną superinteligencją odciąga naszą uwagę od realnych, „cichych" szkód algorytmicznych, które dzieją się tu i teraz. Poprzez analizę zjawisk takich jak stronniczość systemowa (bias), pętle sprzężeń zwrotnych w rekomendacjach czy błędy w biometrii, tekst dowodzi, że techniczna poprawność modelu nie jest tożsama ze sprawiedliwością społeczną. Głównym postulatem jest przejście od pasywnego wdrażania narzędzi ku paradygmatowi „wychowania AI". Oznacza to uznanie systemów za elementy złożonych procesów socjotechnicznych, w których odpowiedzialność nie może rozmyć się między programistą a operatorem. W przeciwieństwie do dystopijnych wizji buntu maszyn, prawdziwym zagrożeniem jest stopniowa erozja ludzkiej sprawczości i zastępowanie krytycznego rozumienia rzeczywistości bezkrytycznym zaufaniem do statystycznej reprezentacji świata.

SŁOWA KLUCZOWE

cicha stronniczość | etyka odpowiedzialności | Raising AI | algorytmiczna niesprawiedliwość | pętle sprzężeń zwrotnych | zmienne zastępcze | proxy variables | sprawiedliwość algorytmiczna | bias statystyczny | efekt Mateusza | automatyzacja decyzji | ryzyko algorytmiczne | AGI | wychowanie sztucznej inteligencji

Asymetria między mocą obliczeniową a odpowiedzialnością

Duchy w maszynie, których nie słychać. Społeczna anatomia algorytmicznej niesprawiedliwości

Rozwój sztucznej inteligencji wprowadził współczesną cywilizację w szczególny stan poznawczego rozdwojenia. Z jednej strony społeczeństwa dysponują coraz doskonalszymi narzędziami przetwarzania informacji, rozpoznawania obrazów, modelowania języka naturalnego, przewidywania zjawisk i wspomagania decyzji; z drugiej zaś nie wykształciły równie zaawansowanych mechanizmów społecznego rozumienia, nadzorowania i oceniania konsekwencji

ich działania. Zdolność technologiczna wyprzedza dojrzałość instytucjonalną, skuteczność obliczeniowa bywa mylona z racjonalnością praktyczną, a statystyczna trafność prognozy uzyskuje niekiedy rangę obiektywnej prawdy o człowieku. Powstaje w ten sposób asymetria charakterystyczna dla przełomowych transformacji technicznych: człowiek potrafi wytwarzać coraz bardziej złożone instrumenty oddziaływania na rzeczywistość, zanim zdoła należycie rozpoznać pełne spektrum ich społecznych następstw. Problem nie sprowadza się do zwyczajnego opóźnienia legislacyjnego wobec innowacji.

Dotyczy on rozdźwięku między mocą obliczeniową a zdolnością formułowania uzasadnionych sądów, między poznaniem probabilistycznym a odpowiedzialnością moralną oraz między efektywnością organizacyjną a sprawiedliwością społeczną. Szczególnie interesującym świadectwem tego stanu jest współczesna wyobraźnia zbiorowa, która nauczyła się obawiać przede wszystkim przyszłej, samoświadomej i potencjalnie wrogiej człowiekowi superinteligencji, przywiązując znacznie mniejszą wagę do codziennych szkód powodowanych przez systemy pozbawione jakichkolwiek potwierdzonych cech świadomości. Popularna kultura utrwaliła obrazy autonomicznych maszyn prowadzących wojnę przeciwko ludzkości, komputerów podejmujących decyzję o wyeliminowaniu załogi statku kosmicznego oraz robotów usiłujących podporządkować człowieka własnej interpretacji jego dobra. Tymczasem współczesne systemy algorytmiczne nie muszą posiadać intencji, przekonań, emocji ani ambicji politycznych, aby wpływać na rozkład szans życiowych milionów ludzi. Wystarczy, że ich wyniki zostaną włączone do procedur udzielania kredytów, rekrutowania pracowników, identyfikowania podejrzanych, przyznawania świadczeń, diagnozowania chorób albo selekcjonowania informacji. Niebezpieczeństwo przyjmuje wówczas postać nie spektakularnego buntu maszyn, lecz rutynowego przekształcania niedoskonałych danych w decyzje wyposażone w autorytet instytucji.

Inspirująca tę analizę książka Chiraga Shaha **Raising AI: How Artificial Intelligence Is Shaping Us-and How We Shape It**, opublikowana w 2026 roku, proponuje odejście od prostego przeciwstawienia technologicznej utopii i katastroficznej dystopii. Jej zasadniczą perspektywą jest rozumienie rozwoju sztucznej inteligencji jako procesu współkształtowania systemów technicznych i środowiska społecznego, w którym funkcjonują. Metafora wychowania, stanowiąca oś tej koncepcji, przesuwa punkt ciężkości z abstrakcyjnego pytania o moment uzyskania przez maszyny inteligencji ogólnej ku konkretnym obowiązkom projektantów, badaczy, przedsiębiorstw, użytkowników i instytucji publicznych. Z perspektywy nauk społecznych szczególnie doniosłe okazuje się rozpoznanie, że nieodpowiedzialność technologiczna nie musi wynikać ze złej woli konstruktorów. Może być produktem ubocznym poprawnie działających procedur optymalizacyjnych, niedostatecznej reprezentatywności danych, niewłaściwie zdefiniowanych funkcji celu, presji organizacyjnej oraz instytucjonalnego przyzwyczajenia do

traktowania wyniku algorytmu jako informacji bardziej wiarygodnej niż doświadczenie konkretnego człowieka. W tym kontekście kluczowe staje się dostrzeżenie problemu cichej stronniczości.

W literaturze dotyczącej technologii i społeczeństwa problem ten można rozpatrywać przez pryzmat klasycznej koncepcji niezamierzonych konsekwencji działania Roberta K. Mertona. Socjolog ten wykazał, że działania podejmowane ze względu na określone cele mogą generować skutki, których sprawcy nie przewidywali i których nie zamierzali wywołać.

W przypadku współczesnych systemów sztucznej inteligencji mechanizm ten uzyskuje szczególną intensywność, ponieważ między decyzją projektową a jej społecznym rezultatem występuje wielowarstwowa infrastruktura pośrednicząca. Projektant może dążyć do zwiększenia trafności klasyfikacji, przedsiębiorstwo do redukcji kosztów, operator do przyspieszenia obsługi, a urząd do ujednolicenia praktyki decyzyjnej.

Systemowa szkodliwość AI wykracza poza intencje twórców

Żaden z uczestników tego procesu nie musi deklarować zamiaru dyskryminacji, mimo że końcowy rezultat działania systemu może polegać na systematycznym pogarszaniu sytuacji określonej grupy obywateli. Zależność ta obnaża ograniczenia moralności skoncentrowanej wyłącznie na indywidualnej intencji sprawcy.

Ocena odpowiedzialności wymaga bowiem uwzględnienia przewidywalności ryzyka, obowiązku należytej staranności, możliwości zapobiegania szkodzie oraz rzeczywistego rozkładu kompetencji decyzyjnych. Problematykę tę pogłębia teoria społeczeństwa ryzyka Ulricha Becka, według której rozwinięta nowoczesność coraz częściej konfrontuje się z zagrożeniami wytwarzanymi przez własne osiągnięcia techniczne, gospodarcze i organizacyjne. Zagrożenia te bywają trudne do bezpośredniego zaobserwowania; wymagają specjalistycznego instrumentarium poznawczego i stają się przedmiotem sporów między ekspertami, instytucjami a obywatelami. W przypadku AI pojawia się jednak dodatkowa trudność: te same systemy, które generują ryzyko, dostarczają jednocześnie informacji służących jego identyfikacji.

Jeżeli przedsiębiorstwo bada bezpieczeństwo własnego modelu głównie za pomocą metryk wyznaczonych podczas projektowania, może przeoczyć szkody pozostające poza zakresem przyjętych wskaźników. Powstaje swoisty problem epistemicznego zamknięcia organizacji, polegający na utożsamieniu tego, co instytucja potrafi zmierzyć, z całokształtem tego, co ma

znaczenie dla społecznej oceny jej działalności. Ryzyko algorytmiczne wymaga zatem nie tylko wiedzy inżynieryjnej, ale i niezależnej kontroli społecznej oraz uwzględnienia doświadczeń osób podlegających automatycznym rozstrzygnięciom. W tym kontekście szczególnie użyteczna okazuje się metafora "duchów w maszynie", obecna w rozważaniach Shaha i jego interpretacjach systemów rekomendacyjnych. Autor opisuje rozmowy badacza z Shelly Plimpton podczas współpracy ze Spotify w 2018 roku, dotyczące nieoczekiwanych konsekwencji optymalizacji wielostronnych platform cyfrowych.

Niezależnie od literackiego charakteru samej metafory, jej znaczenie analityczne można zrekonstruować bez przypisywania modelom obliczeniowym cech metafizycznych. "Duch" oznacza w tym ujęciu zachowanie systemu, którego nie zaplanowano jako celu, lecz które wyłoniło się z oddziaływania danych, reguł uczenia, funkcji optymalizacyjnych i środowiska użytkowania. Nie jest to dowód na autonomię poznawczą czy świadomość maszyny, lecz sygnał, że obserwowany rezultat nie może zostać adekwatnie wyjaśniony wyłącznie przez odwołanie do deklarowanej intencji konstruktorów. System może działać zgodnie z technicznymi założeniami, a jednocześnie wytwarzać nieakceptowalne konsekwencje społeczne. Istotę tego zjawiska dobrze ilustrują platformy rekomendacyjne funkcjonujące jako rynki wielostronne. Platforma muzyczna bowiem nie obsługuje wyłącznie słuchaczy, lecz również twórców, dystrybutorów, reklamodawców i właścicieli praw.

Algorytmiczne pętle sprzężeń zwrotnych wytwarzają preferencje

Algorytm maksymalizujący prawdopodobieństwo dalszego odtwarzania utworów może przyznawać pierwszeństwo treściom już popularnym, gdyż dysponuje o nich większą liczbą obserwacji i silniejszymi sygnałami zainteresowania. Utwory mniej znanych wykonawców otrzymują natomiast ograniczoną ekspozycję, co prowadzi do niewielkiej liczby interakcji i dostarcza systemowi pozornego potwierdzenia ich niższej atrakcyjności. Powstaje dodatnie sprzężenie zwrotne: wcześniejsza widoczność zwiększa przyszłą, podczas gdy ograniczony dostęp do uwagi odbiorców staje się samowzmacniającą przeszkodą rozwoju. Algorytm wcale nie musi zawierać instrukcji nakazującej dyskryminowanie debiutujących artystów.

Wystarczy, że jego funkcja celu preferuje bezpośrednio mierzalne zaangażowanie, a system pomija koszty społeczne i rynkowe wynikające z koncentracji ekspozycji. W ekonomii mechanizm ten można interpretować przez teorię kumulacyjnej przewagi oraz procesy zależności od ścieżki rozwojowej. Robert K. Merton, analizując zjawisko określane jako efekt Mateusza, wskazywał na

mechanizmy sprzyjające przyznawaniu większego uznania podmiotom już posiadającym wysoką pozycję. W badaniach nad gospodarką cyfrową analogiczne zależności opisuje się poprzez rosnące korzyści skali, efekty sieciowe i pętle danych. Im częściej produkt jest prezentowany użytkownikom, tym więcej informacji o nim gromadzi platforma; im więcej ma danych, tym sprawniej przewiduje zainteresowanie; a im trafniejsze są prognozy, tym silniejsza motywacja do dalszej ekspozycji. Mechanizm ten zwiększa efektywność krótkoterminowej optymalizacji, lecz może ograniczać różnorodność rynku, utrudniać odkrywanie nowości i osłabiać konkurencję.

Problem sprawiedliwości nie sprowadza się tutaj do równego traktowania wszystkich utworów, gdyż różnią się one rzeczywistym zainteresowaniem odbiorców. Dotyczy raczej warunków uzyskiwania szansy na to zainteresowanie, które system sam współkształtuje poprzez decyzje o kolejności i częstotliwości prezentacji. Podobna prawidłowość występuje w systemach informacyjnych, usługach pośrednictwa pracy, platformach handlowych i cyfrowych ekosystemach reklamy. Dane gromadzone przez system nie są neutralnym obrazem rzeczywistości niezależnej od niego, ponieważ zachowania użytkowników zależą od tego, jakie możliwości zostały im wcześniej przedstawione. W metodologii badań społecznych mówi się o endogeniczności obserwacji, w informatyce o sprzężeniach zwrotnych między modelem a środowiskiem, a w ekonomii informacji o selektywnym ujawnianiu preferencji pod wpływem dostępnych opcji. W rezultacie przedsiębiorstwo może uznać, że algorytm jedynie odkrywa istniejące zainteresowania, podczas gdy w rzeczywistości częściowo je wytwarza. Ma to daleko idące konsekwencje dla rozumienia autonomii konsumenta, pluralizmu kulturowego i konkurencyjności rynkowej.

Jeżeli system rekomendacyjny współtworzy strukturę dostępnych wyborów, jego odpowiedzialność nie kończy się na poprawnym szacowaniu prawdopodobieństwa kliknięcia. Naukowa analiza cichej stronniczości wymaga jednak wcześniejszego uporządkowania samego pojęcia bias. W języku statystyki termin ten może oznaczać systematyczne odchylenie estymatora od wartości estymowanej, natomiast w konstrukcji niektórych modeli parametr określany jako bias term pełni funkcję wyrazu wolnego i nie zawiera oceny moralnej. W uczeniu maszynowym stronniczość może odnosić się do błędu doboru próby, różnic między populacją treningową a użytkownikami, błędnego pomiaru cech albo ograniczeń reprezentacji danych. W naukach społecznych oznacza ona również mechanizmy nierównego traktowania, reprodukcji stereotypów, dyskryminacji pośredniej oraz systematycznego ograniczania możliwości określonych grup. Rozróżnienie tych znaczeń jest niezbędne: nie każdy statystyczny bias stanowi niesprawiedliwość, podobnie jak nie każda społeczna niesprawiedliwość jest następstwem błędu obliczeniowego.

Techniczna poprawność modelu nie gwarantuje sprawiedliwości społecznej

System może prawidłowo prognozować określone zachowania, a mimo to pozostawać normatywnie kontrowersyjny ze względu na rodzaj wykorzystywanych informacji, cele prognozowania lub skutki zastosowania wyniku. W dokumencie **Artificial Intelligence Risk Management Framework** amerykański National Institute of Standards and Technology wyróżnia trzy szerokie kategorie stronniczości: systemową, obliczeniowo-statystyczną oraz poznawczą, związaną z funkcjonowaniem człowieka.

Klasyfikacja ta jest szczególnie użyteczna, gdyż pozwala uniknąć redukowania całego zagadnienia jedynie do składu zbiorów treningowych. Stronniczość systemowa może wynikać z historycznych nierówności, praktyk instytucjonalnych i sposobu organizacji dostępu do dóbr. Stronniczość obliczeniowo-statystyczna pojawia się wskutek błędów próbkowania, niewłaściwie dobranych zmiennych, metod estymacji lub przesunięcia rozkładu danych. Z kolei stronniczość poznawcza dotyczy zarówno konstruktorów modelu, jak i osób interpretujących jego wyniki – zwłaszcza gdy nadmierne zaufanie do technologii osłabia krytyczną ocenę rekomendacji. Te trzy mechanizmy mogą wzajemnie się wzmacniać, tworząc strukturę przyczynową znacznie bardziej złożoną niż popularne wyobrażenie o algorytmie, który po prostu powiela uprzedzenia swoich programistów.

Z perspektywy socjologii instytucjonalnej szczególnego znaczenia nabiera moment, w którym dane historyczne zostają potraktowane jako wzorzec przyszłych decyzji. Dane o zatrudnieniu nie opisują bowiem wyłącznie kompetencji kandydatów, lecz także rezultaty wcześniejszych praktyk rekrutacyjnych. Dane o przyznawaniu kredytów odzwierciedlają nie tylko zdolność do ich spłaty, ale i dawną politykę banków, strukturę własności majątkowej czy dostępność usług finansowych. Podobnie dane dotyczące interwencji policyjnych są częściowo rezultatem rzeczywistych zdarzeń, lecz również geografii patroli, sposobów zgłaszania przestępstw i intensywności kontroli określonych środowisk. Gdy model uczy się na takich zbiorach, może uznać utrwalony rezultat instytucjonalny za właściwość osoby lub grupy. Ryzyko polega wówczas na pomyleniu cech podmiotu z historią jego traktowania, a następnie na przekształceniu tej historii w prognozę mającą uzasadniać dalsze podobne traktowanie.

Solon Barocas i Andrew D. Selbst w pracy **Big Data's Disparate Impact** wykazali, że intensywne wykorzystanie danych w procesach decyzyjnych może reprodukować nierówności nawet bez jawnego posługiwania się cechami chronionymi. Zmienna opisująca miejsce zamieszkania, historię zawodową, strukturę konsumpcji czy sposób korzystania z usług może pozostawać statystycznie związana z sytuacją społeczną i pochodzeniem – cechami, których bezpośrednie wykorzystanie

podlega ograniczeniom prawnym. Usunięcie z formularza informacji o płci lub narodowości nie eliminuje automatycznie dyskryminacji, jeśli pozostałe zmienne pośrednio odtwarzają te właściwości. W terminologii informatycznej mówi się o zmiennych zastępczych (**proxy variables**), natomiast prawo antydyskryminacyjne pozwala analizować ten problem za pomocą pojęcia nieproporcjonalnie niekorzystnego oddziaływania pozornie neutralnych kryteriów. Nie oznacza to, że każda korelacja jest dowodem bezprawnego działania. Oznacza jednak, że poprawność procedury technicznej nie stanowi wystarczającego argumentu na rzecz sprawiedliwości jej społecznych konsekwencji.

Konflikt matematycznych definicji sprawiedliwości algorytmicznej

Istotnego materiału empirycznego dostarcza spór wokół systemu COMPAS, wykorzystywanego w amerykańskim wymiarze sprawiedliwości do szacowania ryzyka recydywy. W 2016 roku dziennikarze ProPublica przeanalizowali dane osób ocenianych za pomocą tego narzędzia w hrabstwie Broward na Florydzie, zestawiając prognozy z późniejszymi zatrzymaniami i zarzutami. Wśród osób, które w badanym dwuletnim okresie nie popełniły przestępstwa (zgodnie z przyjętą operacjonalizacją), około 45% czarnoskórych oskarżonych zakwalifikowano do kategorii podwyższonego ryzyka, podczas gdy wśród białych odsetek ten wynosił około 23%. Jednocześnie w grupie osób, które faktycznie naruszyły prawo, błędna kwalifikacja do kategorii niskiego ryzyka dotyczyła około 48% białych i 28% czarnoskórych.

Wyniki te ujawniły wyraźne zróżnicowanie struktury błędów predykcyjnych między grupami. Nie pozwalają one jednak na uproszczony wniosek, że biali oskarżeni częściej wracali na drogę przestępstwa lub że każda decyzja oparta na COMPAS była dyskryminacyjna. Spór ten jest szczególnie cenny, gdyż ilustruje konflikt między konkurencyjnymi definicjami sprawiedliwości matematycznej.

Producent systemu, firma Northpointe, kładł nacisk na kalibrację ryzyka – właściwość, według której osoby z podobną oceną probabilistyczną powinny wykazywać zbliżoną częstość występowania prognozowanego zdarzenia, niezależnie od przynależności grupowej. Krytycy zwracali natomiast uwagę na nierówne wskaźniki fałszywych alarmów oraz fałszywego uspokojenia, co ma kluczowe znaczenie dla praw osób poddawanych ocenie. Jon Kleinberg, Sendhil Mullainathan i Manish Raghavan wykazali matematycznie, że przy różnych częstościach występowania zdarzenia w poszczególnych populacjach spełnienie określonych intuicyjnie atrakcyjnych kryteriów sprawiedliwości bywa wzajemnie niemożliwe (poza wyjątkami). Oznacza to,

że konstruktorzy systemów predykcyjnych nie zawsze mogą jednocześnie zapewnić identyczną kalibrację, równowagę błędów i jednakowe traktowanie grup według wszystkich metryk. Wybór kryterium staje się zatem decyzją nie tylko matematyczną, lecz także aksjologiczną i instytucjonalną.

Formalna niemożliwość jednoczesnego spełnienia tych warunków nie jest jednak argumentem za rezygnacją ze sprawiedliwości jako celu projektowego. Przeciwnie – ujawnia konieczność jawnego uzasadnienia, które rodzaje błędów są w danej dziedzinie szczególnie niebezpieczne. Dla porównania: w systemie rekomendacji muzycznych fałszywie dodatni wynik oznacza jedynie zaproponowanie słuchaczowi utworu, który go nie zainteresuje; w wymiarze sprawiedliwości koszt takiego błędu jest nieporównywalnie wyższy.

Matematyczna optymalizacja nie zastąpi sprawiedliwości społecznej

W medycynie pominięcie rzeczywistej choroby może prowadzić do nieodwracalnego pogorszenia stanu zdrowia. W wymiarze sprawiedliwości nieuzasadnione zakwalifikowanie osoby jako jednostki wysokiego ryzyka może ograniczyć jej wolność, choć konkretny skutek zależy od tego, jak narzędzia te wykorzystają właściwe organy. Ekonomiczna funkcja kosztu błędu oraz prawna ocena dopuszczalnego ryzyka powinny zatem pozostawać powiązane z charakterem dobra, na które oddziałuje decyzja. Jednakowa wartość błędu w tabeli statystycznej nie oznacza jednakowej wagi szkody w świecie społecznym.

Rozróżnienie to prowadzi do rozważań nad naturą etyczną tych systemów. W filozofii utylitarystycznej, zapoczątkowanej nowożytnie między innymi przez Jeremy'ego Benthama i rozwijanej przez Johna Stuarta Milla, moralna ocena działania wiąże się z jego konsekwencjami dla dobrostanu. Współczesne modele uczenia maszynowego również projektuje się wokół funkcji celu, lecz utożsamienie matematycznej optymalizacji z utylitaryzmem byłoby nieporozumieniem. Funkcja straty modelu zwykle obejmuje jedynie wybrane, formalnie reprezentowane aspekty rzeczywistości; nie stanowi pełnej teorii ludzkiego dobrostanu, a tym bardziej rozstrzygnięcia sporów o godność, wolność czy sprawiedliwość. Nawet gdy system maksymalizuje poprawność przewidywań, otwarte pozostaje pytanie o zasadność samego celu, koszty jego realizacji i prawa jednostek, które ponoszą konsekwencje błędnych wyników.

W tym miejscu przydatna staje się deontologiczna tradycja Immanuela Kanta, podkreślająca znaczenie traktowania człowieka jako celu samego w sobie, a nie wyłącznie środka do realizacji

cudzych zamierzeń. Nie daje ona gotowego algorytmu rozstrzygania konfliktów, ale przypomina, że niektórych aspektów podmiotowości nie należy redukować do parametrów cudzej funkcji użyteczności.

Dla zobrazowania problemu nierównowagi klasowej autor książki przywołuje modelową sytuację organizacji posiłku dla stu uczniów, z których dziesięcioro nie może spożywać glutenu. System oceniany wyłącznie według liczby osób, którym zapewniono standardowy posiłek, może osiągnąć wynik dziewięćdziesięcioprocentowej skuteczności, całkowicie pomijając potrzeby pozostałych uczestników. Przykład ten nie jest empirycznym testem konkretnego modelu AI, lecz eksperymentem dydaktycznym ujawniającym ograniczenia zagregowanych wskaźników efektywności. W statystyce klasyfikacyjnej analogiczny problem występuje wtedy, gdy jedna klasa zdecydowanie dominuje w próbie. Model przewidujący zawsze klasę większościową może wykazywać pozornie wysoką dokładność ogólną, a jednocześnie całkowicie zawodzić w identyfikowaniu klasy mniejszościowej. Dlatego w analizie takich systemów stosuje się między innymi czułość, swoistość, precyzję, miarę F1, zrównoważoną dokładność oraz analizę macierzy pomyłek. Jednak i te wskaźniki nie zastępują rozpoznania konsekwencji błędów, ponieważ metryka uzasadniona w badaniach przesiewowych może być niewłaściwa przy podejmowaniu decyzji o wolności osobistej lub przyznawaniu świadczeń społecznych.

Kolejnego istotnego przykładu dostarczyły badania Joy Buolamwini i Timnit Gebru opublikowane w 2018 roku pod tytułem **Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification**. Autorki przeanalizowały trzy komercyjne systemy klasyfikujące płeć na podstawie obrazów twarzy i wykazały znaczne zróżnicowanie dokładności w zależności od odcienia skóry i płci ocenianych osób. W najbardziej niekorzystnym przypadku wskaźnik błędnej klasyfikacji kobiet o ciemniejszej skórze osiągał 34,7%, podczas gdy maksymalny błąd w grupie mężczyzn o jaśniejszej skórze wynosił zaledwie 0,8%. Badaczki wykazały również dominację osób o jaśniejszym odcieniu skóry w analizowanych popularnych zbiorach referencyjnych, gdzie ich udział wynosił odpowiednio 79,6% oraz 86,2%.

Wyniki te nie uprawniają do automatycznego uogólnienia na każdy system rozpoznawania tożsamości, gdyż przedmiotem eksperymentu była klasyfikacja płci. Stanowią natomiast mocny dowód, że zagregowane wyniki testów mogą ukrywać znaczne dysproporcje w jakości działania między podgrupami, zwłaszcza gdy uwzględnia się jednocześnie kilka cech demograficznych. Badanie **Gender Shades** ma znaczenie wykraczające poza informatyczną analizę obrazów, ponieważ dostarcza empirycznego przykładu intersekcjonalności, rozwijanej w teorii prawa i naukach społecznych między innymi przez Kimberlé Crenshaw. Intersekcjonalność zwraca uwagę, że położenia społecznego jednostki nie można zawsze wyjaśnić przez oddzielne rozpatrywanie

poszczególnych cech lub kategorii przynależności. Niekorzystny rezultat może ujawniać się szczególnie intensywnie na przecięciu wielu wymiarów, choć analiza każdej kategorii osobno nie pokazuje pełnej skali problemu. Przeniesienie tej perspektywy do audytów systemów AI oznacza konieczność badania nie tylko średniej trafności dla kobiet i mężczyzn albo dla poszczególnych odcieni skóry, lecz także jakości działania wobec grup tworzonych przez współwystępowanie tych cech. Wymaga to odpowiednich danych, procedur ochrony prywatności i statystycznie uzasadnionych metod porównawczych. Nie może polegać na dowolnym mnożeniu kategorii bez kontroli liczebności prób, gdyż analiza niewielkich podgrup jest szczególnie podatna na niepewność estymacji.

Zagrożenia związane z automatycznym rozpoznawaniem twarzy zostały następnie zbadane przez amerykański National Institute of Standards and Technology.

Błąd algorytmiczny jako element systemowej opresji

W opublikowanym w 2019 roku obszernym raporcie dotyczącym skutków zróżnicowania demograficznego wykazano, że w wielu analizowanych algorytmach występowały znaczące różnice w częstości fałszywie dodatnich dopasowań między populacjami. W zadaniach weryfikacyjnych porównujących dwie fotografie różnice te osiągały w niektórych systemach rząd dziesięcio-, a nawet stukrotności, choć zdarzały się istotne wyjątki zależne od konstrukcji i pochodzenia algorytmu. Raport nie dowodził, że wszystkie systemy mają identyczne defekty ani że jedyną przyczyną różnic jest skład danych treningowych. Wskazywał natomiast, że pomiar skuteczności rozpoznawania tożsamości powinien uwzględniać sposób wykorzystania systemu, rodzaj błędu, charakter danych oraz populację, wobec której technologia ma być stosowana. Z perspektywy praw jednostki szczególnego znaczenia nabiera sytuacja, w której fałszywe dopasowanie przestaje być techniczną anomalią, a staje się początkiem czynności organów ścigania.

Takie następstwo wystąpiło w sprawie Roberta Williamsa, niesłusznie zatrzymanego przez policję w Detroit w styczniu 2020 roku. W toku postępowania wykorzystano błędne dopasowanie obrazu twarzy z monitoringu, a następnie zastosowano dalsze czynności identyfikacyjne obciążone wadami proceduralnymi. Williams został pozbawiony wolności mimo braku rzeczywistego związku z zarzucanym czynem. W 2024 roku zawarto ugodę, która doprowadziła do wprowadzenia istotnych ograniczeń w policyjnym stosowaniu technologii rozpoznawania twarzy, w tym wymogu posiadania niezależnych i wiarygodnych dowodów przed dokonaniem zatrzymania.

Ten przypadek pokazuje, że szkoda nie musi być wyłącznym następstwem błędu modelu. Powstaje ona na styku niedoskonałej identyfikacji, decyzji funkcjonariuszy, procedur dowodowych i

niewystarczającego nadzoru instytucjonalnego. Odpowiedzialność nie znika zatem w momencie włączenia człowieka do procesu decyzyjnego, gdyż sam fakt ludzkiego udziału nie gwarantuje krytycznej oceny wyniku algorytmicznego. Przypadek ten można osadzić w szerszym sporze o społeczne znaczenie nadzoru i mechanizmów dyscyplinowania. Michel Foucault analizował przemiany nowoczesnych instytucji, wykazując, że władza nie musi działać wyłącznie poprzez spektakularny przymus, lecz może funkcjonować dzięki rozproszonym technikom obserwacji, klasyfikacji, normalizacji i kontroli. Współczesne systemy analizy obrazu, identyfikacji biometrycznej oraz profilowania predykcyjnego nie są prostą kontynuacją opisywanych przez niego instytucji, ale umożliwiają ponowne postawienie pytań o asymetrię widzialności i zakres kontroli nad jednostką. Obywatel może być stale poddawany obserwacji bez wiedzy o sposobie przetwarzania danych i kryteriach przypisywania mu określonych cech. Zarazem instytucja może wykorzystywać wyniki systemu, nie dysponując pełną wiedzą o jego wewnętrznych zależnościach statystycznych. Dochodzi wówczas do szczególnego odwrócenia relacji poznawczej: władza dysponuje coraz większą ilością danych o jednostkach, ale niekoniecznie coraz lepszym rozumieniem przyczyn własnych decyzji.

Cicha stronniczość nie ogranicza się jednak do sytuacji, w których algorytm przypisuje człowiekowi ryzyko przestępczości albo błędnie rozpoznaje jego twarz. Safiya Umoja Noble w monografii **Algorithms of Oppression: How Search Engines Reinforce Racism** wykazała, że komercyjne systemy wyszukiwania informacji również mogą reprodukować i wzmacniać społecznie szkodliwe stereotypy. Jej analiza dotyczyła między innymi historycznych wyników wyszukiwania związanych z kobietami należącymi do mniejszości rasowych oraz sposobów seksualizowania tych kategorii w sieci. Doniosłość tego problemu wynika z faktu, że przeciętny użytkownik nie postrzega wyszukiwarki jako redaktora dokonującego normatywnej selekcji treści. Interfejs przedstawia wyniki w formie technicznie uporządkowanej, co sugeruje bezstronność, podczas gdy ranking może pozostawać produktem komercyjnych zależności, historycznej struktury zasobów internetowych, optymalizacji trafności oraz zachowań zbiorowych. Użytkownik nierzadko otrzymuje więc nie neutralne odzwierciedlenie wiedzy społecznej, lecz wynik wielostronnego procesu selekcji, którego kryteriów nie zna.

Niewidzialność szkód algorytmicznych i koszty społeczne automatyzacji

Mechanizm ten można wyjaśnić za pomocą teorii władzy symbolicznej Pierre'a Bourdieu. Władza ta oddziałuje nie tylko poprzez bezpośrednie nakazy, lecz przede wszystkim przez ustanawianie

kategorii rozpoznawania rzeczywistości — określanie tego, co uznajemy za ważne, wiarygodne, oczywiste lub godne uwagi. Wyszukiwarki i systemy rekomendacyjne uczestniczą w tych procesach, organizując dostęp do informacji i współkształtując hierarchię widoczności podmiotów. Nie oznacza to jednak, że każdy ranking jest instrumentem świadomej manipulacji. Problem jest subtelniejszy: techniczny porządek prezentacji może zyskać status porządku poznawczego w momencie, gdy odbiorca przestaje odróżniać wynik procesu selekcji od rzeczywistej struktury świata. Ma to kluczowe znaczenie dla pluralizmu debaty publicznej, widoczności mniejszości, konkurencji gospodarczej oraz reprodukcji stereotypów kulturowych.

Z perspektywy teorii organizacji pojawia się ponadto ryzyko instytucjonalnej normalizacji odstępstw. Diane Vaughan, analizując mechanizmy, które doprowadziły do katastrofy promu Challenger, opisała proces, w którym powtarzające się nieprawidłowości przestają być postrzegane jako sygnały zagrożenia, o ile ich wcześniejsze wystąpienia nie wywołały bezpośredniej katastrofy. Analogiczna logika może przejawiać się w zarządzaniu systemami AI: kolejne błędne rekomendacje, fałszywe dopasowania czy nierówności wskaźników jakości bywają traktowane jako dopuszczalny koszt funkcjonowania infrastruktury. O ile awaria urządzenia fizycznego ujawnia się poprzez wyraźne naruszenie parametrów technicznych, o tyle szkoda algorytmiczna bywa rozproszona, odroczona i trudna do przypisania konkretnemu zdarzeniu.

Osoba odrzucona w rekrutacji nie musi wiedzieć, że przyczyną była wadliwa klasyfikacja; przedsiębiorca może nie poznać powodów ograniczenia widoczności swojej oferty; obywatel może otrzymać decyzję administracyjną, nie mając pojęcia, jaką rolę odegrała automatyczna ocena ryzyka. Ta niedostateczna obserwowalność skutków sprzyja ich organizacyjnemu lekceważeniu. Odrębnego rozpoznania wymaga związek stronniczości algorytmicznej z ekonomią zachowań organizacyjnych. Każde przedsiębiorstwo funkcjonuje w systemie bodźców obejmującym koszty kapitału, oczekiwania inwestorów, presję konkurencyjną czy strukturę wynagrodzeń kierownictwa. Jeśli sukces wdrożenia AI mierzy się niemal wyłącznie redukcją czasu obsługi, ograniczeniem zatrudnienia lub liczbą przetworzonych spraw, organizacja może systematycznie niedoszacowywać kosztów błędnych decyzji, które są ponoszone poza jej własnym rachunkiem wyników.

W ekonomii dobrobytu zjawisko to wiąże się z efektami zewnętrznymi — konsekwencjami działalności gospodarczej nieuwzględnionymi w mechanizmie cenowym. Automatyzacja może zatem zwiększać produktywność firmy, jednocześnie przenosząc część kosztów na konsumentów, pracowników, organy odwoławcze, służby publiczne lub całe grupy społeczne. Rzetelna ocena efektywności technologii wymaga więc uwzględnienia kosztu społecznego, a nie jedynie oszczędności finansowych podmiotu wdrażającego system.

AI jako proces socjotechniczny wymaga realnej odpowiedzialności instytucjonalnej

W tym miejscu metafora „goblinów", rozwijana w książce obok pojęcia „duchów", zyskuje odrębną funkcję analityczną. O ile „duchy" oznaczają przede wszystkim niezamierzone i trudne do przewidzenia właściwości złożonych systemów, o tyle „gobliny" symbolizują błędy wynikające z zaniedbań, pośpiechu, złej organizacji prac czy świadomego pomijania wymogów należytej staranności. Rozróżnienie to jest użyteczne, choć nie należy nadawać mu charakteru ścisłej klasyfikacji naukowej.

W praktyce oba mechanizmy mogą współwystępować: organizacja ignorująca testy odporności i audyty danych zwiększa ryzyko, że nieoczekiwane zachowania systemu ujawnią się dopiero po wyrządzeniu szkody. Nieprzewidywalność wynikająca ze złożoności modelu nie zdejmuje jednak obowiązku przewidywania typowych kategorii ryzyka. Im dotkliwsze mogą być skutki błędnej decyzji, tym silniejsze powinny być wymagania dotyczące jakości danych, dokumentacji projektu, niezależnej oceny, monitorowania efektów oraz możliwości ograniczenia lub wycofania systemu z eksploatacji. Analiza Andrew D. Selbst i współautorów, przedstawiona w pracy **Fairness and Abstraction in Sociotechnical Systems**, wskazuje na jeszcze głębszą trudność. Projektowanie sprawiedliwego algorytmu wymaga nie tylko prawidłowego wyboru matematycznej miary sprawiedliwości, lecz także uniknięcia nadmiernej abstrakcji od społecznego środowiska jego zastosowania. Formalne przedstawienie problemu decyzyjnego może pomijać relacje władzy, instytucjonalne motywacje, historyczne uwarunkowania danych, możliwości odwołania się od rozstrzygnięcia czy zachowania osób objętych działaniem systemu. Algorytm może spełniać określone kryterium uczciwości w warunkach laboratoryjnych, a mimo to w rzeczywistości wytwarzać szkodliwe rezultaty. Zdolność do minimalizowania wskazanej funkcji straty nie dowodzi bowiem, że funkcja ta prawidłowo reprezentuje społecznie istotny problem.

Właśnie dlatego systemy AI powinny być badane jako elementy procesów socjotechnicznych, a nie wyłącznie jako samodzielne obiekty matematyczne. Konsekwencje przyjęcia tej perspektywy są szczególnie istotne dla nauk prawnych. W tradycji państwa prawnego decyzja wpływająca na sytuację jednostki wymaga nie tylko odpowiedniej podstawy kompetencyjnej, lecz także poszanowania standardów proceduralnych – w tym możliwości poznania przesłanek rozstrzygnięcia, zakwestionowania błędu i skorzystania z właściwych środków ochrony. System algorytmiczny może zostać włączony do procesu decyzyjnego tak, by formalnie zachować udział urzędnika lub pracownika, ale faktycznie ograniczyć jego niezależność poznawczą. Jeżeli osoba odpowiedzialna za końcowe rozstrzygnięcie nie rozumie zasad działania systemu, nie ma dostępu

do danych umożliwiających weryfikację wyniku albo podlega silnej presji organizacyjnej, by akceptować automatyczne rekomendacje, nadzór człowieka staje się pozorny. W takich warunkach deklarowana zasada ludzkiej kontroli nie gwarantuje rzeczywistej sprawczości. Konieczne jest więc rozróżnienie obecności człowieka w procedurze od jego faktycznej zdolności do zrozumienia, zakwestionowania i zmiany decyzji proponowanej przez system. Problem ma również wymiar epistemologiczny: w klasycznej tradycji filozoficznej wiedza wymaga nie tylko sformułowania sądu, lecz także odpowiedniego uzasadnienia jego prawdziwości.

W zastosowaniach AI często dochodzi do praktycznego rozdzielenia skuteczności prognozy i możliwości przedstawienia pełnego wyjaśnienia mechanizmu prowadzącego do wyniku. Nie oznacza to, że każdy model uczenia maszynowego jest nieinterpretowalny ani że każda decyzja statystyczna stanowi arbitralny akt władzy. Oznacza jednak, że różne formy wyjaśnialności odpowiadają na odmienne potrzeby: inżynier oczekuje informacji o znaczeniu zmiennych, audytor o źródłach błędu, sędzia o związku wyniku z przesłankami prawnymi, a obywatel o powodach niekorzystnej decyzji i możliwościach jej zakwestionowania. Właściwe uzasadnienie powinno więc odpowiadać zarówno na pytanie o mechanizm obliczeniowy, jak i o dopuszczalność wykorzystania tego mechanizmu w określonym kontekście społecznym.

Zasadniczym zagrożeniem dla odpowiedzialności staje się tu swoiste rozproszenie autorstwa decyzji. Inżynier opracowujący model odpowiada za konstrukcję systemu, dostawca infrastruktury za parametry usługi, podmiot wdrażający za organizację procesu, operator za sposób wykorzystania wyniku, a instytucja publiczna za legalność swojego działania. Jeżeli żadna z tych kategorii odpowiedzialności nie zostanie jasno określona, może powstać sytuacja, w której każdy uczestnik procesu wskazuje na innego jako właściwego adresata roszczenia lub krytyki. Klasyczne zagadnienie wielu rąk, analizowane w filozofii administracji i teorii odpowiedzialności organizacyjnej, nabiera w środowisku algorytmicznym nowego znaczenia. Złożoność techniczna może utrudniać przypisanie przyczynowości, ale nie powinna sama w sobie prowadzić do immunizacji instytucji przed odpowiedzialnością za skutki stosowanych przez nie narzędzi.

Negatywne następstwa automatyzacji nie zawsze mają postać naruszenia praw jednostkowych, które można bezpośrednio wykazać w pojedynczej sprawie. Część szkód ma charakter zbiorowy i kumulacyjny. Stopniowe zmniejszanie widoczności określonych środowisk kulturowych, ograniczanie różnorodności informacji, pogłębianie zależności obywateli od zamkniętych platform, utrwalanie regionalnych nierówności technologicznych czy redukowanie możliwości zdobywania doświadczenia zawodowego przez początkujących pracowników mogą oddziaływać na całą strukturę społeczną.

Wielopoziomowe zarządzanie ryzykiem między aktualnymi szkodami a przyszłą superinteligencją

W analizie tych procesów szczególnie użyteczna okazuje się perspektywa dóbr wspólnych i infrastruktury społecznej. Niektóre zasoby – zaufanie publiczne, pluralizm informacyjny, kompetencje poznawcze społeczeństwa czy instytucjonalna zdolność do niezależnej kontroli – nie dają się łatwo sprowadzić do sumy indywidualnych preferencji konsumenckich.

Ich osłabienie może następować stopniowo, nawet wtedy, gdy poszczególni użytkownicy uznają konkretne usługi AI za wygodne i użyteczne. Należy jednak unikać błędnego utożsamiania sztucznej inteligencji ze źródłem wszelkich patologii społecznych. Dyskryminacja, arbitralność administracyjna, asymetria informacyjna, koncentracja gospodarcza czy nierówny dostęp do dóbr istniały na długo przed rozwojem uczenia maszynowego. Systemy AI mogą pewne problemy intensyfikować, ale w określonych zastosowaniach potrafią również ograniczać wpływ ludzkich uprzedzeń, poprawiać spójność procedur, zwiększać dostępność usług i ujawniać nierówności, które wcześniej pozostawały poza zasięgiem analizy. Kluczowa jest zatem ocena porównawcza: rzeczywiste rezultaty systemu algorytmicznego należy zestawiać z realistycznym sposobem podejmowania decyzji bez jego zastosowania, a nie z idealizowanym obrazem nieomylnego i bezstronnego człowieka. Wdrożenie AI może być uzasadnione, jeśli wiarygodne badania wykazują ograniczenie liczby szkód, zwiększenie możliwości odwoławczych oraz poprawę dostępności świadczeń. Samo występowanie błędów nie przesądza o niedopuszczalności technologii; rozstrzygający wpływ ma ich charakter, skala, odwracalność, rozkład społeczny oraz skuteczność mechanizmów zapobiegawczych.

Podobnej ostrożności wymaga interpretacja zagrożeń związanych z przyszłą sztuczną inteligencją ogólną. Skupienie uwagi na aktualnych szkodach algorytmicznych nie dowodzi, że ryzyka wynikające z potencjalnych systemów o znacznie większej autonomii są pozbawione znaczenia. Obecny stan wiedzy nie pozwala na wiarygodne wyznaczenie daty powstania AGI ani na jednoznaczne określenie wszystkich konsekwencji jej ewentualnego osiągnięcia. Nie oznacza to jednak, że bieżące problemy społeczne należy odkładać do czasu rozwiązania wielkich zagadnień przyszłej superinteligencji. Właściwsza jest perspektywa wielopoziomowego zarządzania ryzykiem, obejmująca jednocześnie szkody występujące już dziś, nowe zagrożenia wynikające ze wzrostu sprawczości systemów oraz niepewności dotyczące dalszych kierunków rozwoju technologicznego. Dobrze zaprojektowane instytucje powinny potrafić reagować zarówno na rozpoznane następstwa obecnych wdrożeń, jak i na zagrożenia, których prawdopodobieństwo pozostaje przedmiotem badań.

Wychowanie AI jako ochrona ludzkiej sprawczości

W ten sposób dochodzimy do istoty proponowanego paradygmatu wychowania sztucznej inteligencji. Nie polega on na przypisywaniu algorytmom dziecięcej psychiki, ludzkich emocji czy moralnej niedojrzałości w dosłownym znaczeniu. Metafora wychowania staje się użyteczna dopiero wtedy, gdy przełożymy ją na konkretne obowiązki człowieka i instytucji: rozpoznawanie możliwości oraz ograniczeń systemu, kształtowanie warunków jego działania, stopniowanie kompetencji, konfrontowanie wyników z rzeczywistością, monitorowanie skutków społecznych i zachowywanie możliwości skutecznej interwencji. O ile wychowanie w sensie pedagogicznym ma za przedmiot rozwój osoby, o tyle proces dostrajania sztucznej inteligencji dotyczy technologii, której status ontologiczny i moralny pozostaje odrębnym zagadnieniem. Wspólnym mianownikiem obu perspektyw jest odpowiedzialność za warunki rozwoju i skutki udzielania autonomii, jednak analogia ta nie znosi zasadniczych różnic między ludzkim uczeniem się a obliczeniową optymalizacją parametrów.

Współczesna cywilizacja staje przed potrzebą przeformułowania samego pojęcia postępu technicznego. Tradycyjna narracja modernizacyjna często zakładała, że wzrost wydajności, automatyzacja pracy i rozszerzanie zdolności poznawczych prowadzą w dłuższej perspektywie do poprawy warunków życia. Historia gospodarcza dostarcza licznych argumentów na rzecz tej zależności, ale wskazuje również, że korzyści z przełomowych technologii nie rozkładają się automatycznie i równomiernie. Ich społeczna dystrybucja zależy od instytucji rynku pracy, struktur własności, regulacji prawnych, systemów edukacji oraz politycznej zdolności do korygowania niepożądanych następstw. Sztuczna inteligencja potęguje ten problem, stając się technologią ogólnego zastosowania, która wpływa nie tylko na fizyczne wykonywanie zadań, lecz także na wytwarzanie informacji, ocenę rzeczywistości i procesy podejmowania decyzji.

W takim środowisku ochrona ludzkiej sprawczości przestaje być wyłącznie postulatem filozoficznym, a staje się warunkiem trwałości instytucji społecznych. Rozpatrywana z tej perspektywy cicha stronniczość jest czymś więcej niż zbiorem błędów statystycznych, które powinna naprawić kolejna generacja modeli. Stanowi ona objaw szerszego problemu cywilizacyjnego: rozchodzenia się zdolności przetwarzania informacji i zdolności rozumienia skutków ich wykorzystania. System może poprawnie wykonywać operacje matematyczne, a jednocześnie uczestniczyć w niesprawiedliwym procesie społecznym. Instytucja może zwiększać wydajność administracyjną, równocześnie pogarszając zdolność obywatela do dochodzenia swoich praw. Przedsiębiorstwo może doskonalić personalizację usług, a zarazem wzmacniać zależność użytkowników od mechanizmów selekcji, których nie potrafią samodzielnie ocenić. Odpowiedzialny rozwój AI wymaga zatem wyjścia poza technologiczną definicję sukcesu ku modelowi, w którym

efektywność, bezpieczeństwo, sprawiedliwość i autonomia człowieka podlegają równoczesnej – choć nie zawsze pozbawionej konfliktów – ocenie.

Refleksja ta pozwala odczytać w nowym kontekście intuicję Thomasa Stearnsa Eliota, według której kres określonego porządku nie musi przyjmować postaci widowiskowej katastrofy, lecz może następować przez stopniową utratę jego wewnętrznej zdolności do trwania. W odniesieniu do społeczeństwa cyfrowego oznacza to konieczność dostrzegania zjawisk, które nie wywołują natychmiastowej sensacji, choć długofalowo przekształcają strukturę ludzkich możliwości.

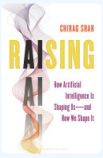
Człowiek nie musi zostać podbity przez inteligentną maszynę, aby utracić część sprawczości; wystarczy, że kolejne instytucje przestaną uzasadniać swoje decyzje inaczej niż poprzez odwołanie do wyników systemów, których nie potrafią krytycznie oceniać. Nie musi również zostać pozbawiony formalnych praw, by ich realizacja stała się trudniejsza wskutek automatycznych procedur, asymetrii informacyjnych i niewidocznych mechanizmów selekcji. Analiza cichej stronniczości prowadzi tym samym do rozpoznania zasadniczej sprzeczności epoki algorytmicznej: im skuteczniejsze stają się systemy służące porządkowaniu rzeczywistości, tym silniejsza pojawia się pokusa utożsamiania wytworzonego przez nie porządku z samą rzeczywistością.

Historia nauki uczy jednak, że nawet najbardziej użyteczne modele pozostają selektywnymi reprezentacjami świata, zależnymi od przyjętych założeń, dostępnych danych i celów poznawczych. Przejście od statystycznej reprezentacji do decyzji dotyczącej ludzkiego życia wymaga dodatkowego uzasadnienia, którego nie zastąpi ani wysoka dokładność obliczeń, ani prestiż firmy technologicznej, ani atrakcyjność interfejsu. Rozpoznanie tej granicy jest warunkiem rozwoju sztucznej inteligencji jako narzędzia wzmacniającego ludzkie możliwości, a nie stopniowo zastępującego ludzką odpowiedzialność. Właśnie tutaj problem rzeczywistych szkód technologicznych spotyka się z wyobrażeniami o przyszłej AI. Opowieści o HAL-u 9000, Skynetcie i autonomicznych robotach nie są wyłącznie fantazjami na temat maszyn, które jeszcze nie istnieją.

Stanowią one kulturowe eksperymenty myślowe, za pomocą których społeczeństwa usiłują określić granice dopuszczalnej autonomii technicznej, naturę odpowiedzialności konstruktorów i sens zachowania ludzkiej kontroli. Ich wartość nie polega na wiarygodnym przewidywaniu przyszłości, lecz na zdolności ujawniania ukrytych założeń współczesnej kultury technologicznej. Zderzenie tych narracji z wiedzą naukową pozwala zrozumieć, dlaczego cywilizacja od dziesięcioleci wyobraża sobie inteligentne maszyny jako potencjalnych następców człowieka, a znacznie rzadziej jako zwierciadła jego własnych uprzedzeń, aspiracji, konfliktów interesów i instytucjonalnych niedoskonałości.

> Czy lęk przed autonomicznym buntem maszyn nie jest jedynie wygodną zasłoną, która pozwala nam ignorować fakt, że już teraz stajemy się zakładnikami własnych uprzedzeń zaklętych w kodzie? Prawdziwa katastrofa może nie nadejść wraz z przebudzeniem superinteligencji, lecz z momentu, w którym przestaniemy odróżniać matematyczny model świata od samej rzeczywistości. W tym scenariuszu człowiek nie zostaje podbity przez maszynę, lecz dobrowolnie rezygnuje z odpowiedzialności za własne decyzje.

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 oraz tytuł magistra informatyki na Uniwersytecie Massachusetts w Amherst i Indyjskim Instytucie Technologii w Madrasie. Shah jest dyrektorem założycielem InfoSeeking Lab i współdyrektorem RAISE (Centrum Odpowiedzialności w Systemach i Doświadczeniach AI). Jego badania koncentrują się na inteligentnym dostępie do informacji, zorientowanych na człowieka i zadania systemów wyszukiwania i rekomendacji, wspólnym wyszukiwaniu informacji oraz odpowiedzialnej sztucznej inteligencji. Jako wybitny członek ACM i ASIS&T, prowadził szeroko zakrojone badania nad uczciwością, przejrzystością i etyką w uczeniu maszynowym i generatywnych systemach sztucznej inteligencji.

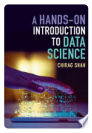
***Chirag Shah** is a Professor in the Information School and an Adjunct Professor in the Paul G. Allen School of Computer Science & Engineering at the University of Washington. He received his Ph.D. in Information Science from the University of North Carolina at Chapel Hill and master's degrees in computer science from the University of Massachusetts Amherst and the Indian Institute of Technology Madras. Shah is the founding director of the InfoSeeking Lab and co-director of RAISE (Center for Responsibility in AI Systems & Experiences). His research centers on intelligent information access, human-centered and task-oriented search and recommendation systems, collaborative information retrieval, and responsible AI. A Distinguished Member of the ACM and ASIS&T, he has extensively investigated fairness, transparency, and ethics in machine learning and generative artificial intelligence systems.*

University of Washington

Literatura uzupełniająca

Książki

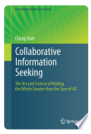
Książki autora i dzieła przywołane w tekście:



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

Chirag Shah

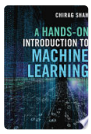
2020 | Cambridge University Press | ISBN: 9781108472449



[2] "Collaborative Information Seeking"

Chirag Shah

2012 | Springer Science & Business Media | ISBN: 9783642288135



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

Chirag Shah

2022 | Cambridge University Press | ISBN: 9781009123303



[4] "Task Intelligence for Search and Recommendation"

Chirag Shah, Ryen W. White

2022 | Springer Nature | ISBN: 9783031023262



[5] "Agent Nation"

Chirag Shah

2026 | Springer Nature | ISBN: 9798868824548



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

Jiqun Liu, Chirag Shah

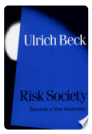
2022 | Springer Nature | ISBN: 9783031023194

Książki autorów przywoływanych i literatura pokrewna:



[7] "Social Theory and Social Structure"

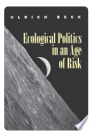
Robert K. Merton



[8] "Risk Society"

Ulrich Beck

SAGE | ISBN: 9781446223420



[9] "Ecological Politics in an Age of Risk"

Ulrich Beck

2018 | John Wiley & Sons | ISBN: 9780745692678



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

Sendhil Mullainathan

Penguin UK | ISBN: 9780141961194



[11] **"Introduction to the Principles of Morals and Legislation"**

Jeremy Bentham

Good Press

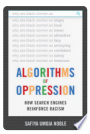
[12] **"A Protest against Law-Taxes"**

Jeremy Bentham



[13] **"Panopticon or the Inspection-House"**

Jeremy Bentham



[14] **"Algorithms of Oppression"**

Safiya Umoja Noble

2018 | NYU Press | ISBN: 9781479837243

[15] **"Facing the Coded Gaze with Evocative Audits and Algorithmic Audits"**

Joy Buolamwini

2022



[16] **"Black in AI"**

Timnit Gebru

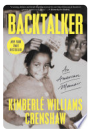
Berrett-Koehler Publishers | ISBN: 9798890571908



[17] **"Demarginalizing the Intersection of Race and Sex"**

Kimberlé Crenshaw

GRIN Verlag | ISBN: 9783346202895



[18] **"Backtalker"**

Kimberlé Crenshaw

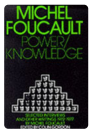
Simon and Schuster | ISBN: 9781982181000



[19] **"Power"**

Michel Foucault

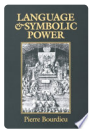
2019 | Penguin UK | ISBN: 9780141991375



[20] **"Power/knowledge"**

Michel Foucault

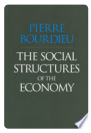
1980 | Longman | ISBN: 9780710800718



[21] **"Language and symbolic power"**

Pierre Bourdieu

Harvard University Press | ISBN: 9780674510418



[22] **"The Social Structures of the Economy"**

Pierre Bourdieu

2005 | Polity | ISBN: 9780745625393

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] **"Individualization: Institutionalized Individualism and its Social and Political Consequences"**

Ulrich Beck, Elisabeth Beck-Gernsheim

2002 | SAGE Publications Ltd eBooks | DOI: 10.4135/9781446218693

Źródło

[2] **"Risk Society, Towards a New Modernity"**

William Leiss, Ulrich Beck, Mark Ritter, Scott M. Lash, Brian Wynne

1994 | The Canadian Journal of Sociology | DOI: 10.2307/3341155

Źródło

[3] **"Unpacking cosmopolitanism for the social sciences: a research agenda"**

Ulrich Beck, Natan Sznaider

2006 | British Journal of Sociology | DOI: 10.1111/j.1468-4446.2006.00091.x

Źródło

[4] **"Living in the world risk society: A Hobhouse Memorial Public Lecture given on Wednesday 15 February 2006 at the London School of Economics"**

Ulrich Beck

2006 | Economy and Society | DOI: 10.1080/03085140600844902

Źródło

[5] "Language (Technology) is Power: A Critical Survey of “Bias” in NLP"

Su Lin Blodgett, Solon Barocas, Hal Daumé, Hanna Wallach

2020 | Proceedings of the conference - Association for Computational Linguistics. Meeting | DOI: 10.18653/v1/2020.acl-main.485

Źródło

[6] "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

[7] "Algorithmic Fairness: Choices, Assumptions, and Definitions"

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D’Amour, Kristian Lum

2020 | Annual Review of Statistics and Its Application | DOI: 10.1146/annurev-statistics-042720-125902

Źródło

[8] "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

[9] "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

[10] "An Institutional View Of Algorithmic Impact Assessments"

Andrew D. Selbst

2021 | SSRN Electronic Journal

Źródło

[11] "Negligence and AI's Human Users"

Andrew D. Selbst

2019 | SSRN Electronic Journal

Źródło

[12] "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

[13] "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

[14] "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

[15] "Group formation in large social networks: membership, growth, and evolution"

Lars Bäckström, D.P. Huttenlocher, Jon Michael Kleinberg, Xiangyang Lan

2006 | ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) | DOI: 10.1145/1150402.1150412

Źródło

[16] "Enjoying the Quiet Life? Corporate Governance and Managerial Preferences"

Marianne Bertrand, Sendhil Mullainathan

2003 | Journal of Political Economy | DOI: 10.1086/376950

Źródło

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

Jon Kleinberg, Sendhil Mullainathan

2019 | DOI: 10.3386/w25854

Źródło

[18] "The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization"

Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan

2023 | Management Science | DOI: 10.1287/mnsc.2022.03683

Źródło

[19] "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

[20] "Sistema de la ciencia social : ideado por el jurisconsulto inglés Jeremias Bentham"

Toribio Núñez, Jeremy Bentham

1820 | Hispana

Źródło

[21] "Principles of Political Economy: With Some of Their Applications to Social Philosophy"

John Stuart Mill

1848

Źródło

[22] "3334 FROM JOHN STUART MILL Between 18 and 24 April 1827"

John Stuart Mill

2006 | The Collected Works of Jeremy Bentham: The Correspondence of Jeremy Bentham, Vol. 12: July 1824 to June 1828 | DOI: 10.1093/oseo/instance.00067625

Źródło

[23] "Will having fun make you happier than studying?"

J. Mill

2021 | Philosophy for Kids | DOI: 10.4324/9781003237150-8

Źródło

[24] "Critique of the Power of Judgment"

Immanuel Kant, Eric Matthews

2000 | Cambridge University Press eBooks | DOI: 10.1017/cbo9780511804656

Źródło

[25] "Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem"

Sasha Costanza-Chock, Inioluwa Deborah Raji, Joy Buolamwini

2022 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3531146.3533213

Źródło

[26] "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜"

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell

2021 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3442188.3445922

Źródło

[27] "Model Cards for Model Reporting"

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Timnit Gebru

2019 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3287560.3287596

Źródło

[28] "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing"

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru

2020 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3351095.3372873

Źródło

[29] "Fine-Grained Recognition in the Wild: A Multi-task Domain Adaptation Approach"

Timnit Gebru, Judy Hoffman, Li Fei-Fei

2017 | IEEE/CVF International Conference on Computer Vision (ICCV) | DOI: 10.1109/iccv.2017.151

Źródło

[30] "From Private Violence to Mass Incarceration: Thinking Intersectionally about Women, Race, and Social Control"

Kimberle Williams Crenshaw

2012 | Scholarship Archive (Columbia Law School)

Źródło

[31] "Gender imbalance in infant mortality: A cross-national study of social structure and female infanticide"

Kana Fuse, Edward M. Crenshaw

2006 | Social Science & Medicine | DOI: 10.1016/j.socscimed.2005.06.006

Źródło

[32] "Power/Knowledge: Selected Interviews and Other Writings 1972-1977"

Michel Foucault, Colin Gordon

1980

Źródło

[33] "The Subject and Power"

Michel Foucault

1982 | Critical Inquiry | DOI: 10.1086/448181

Źródło

[34] "Human Nature: Justice vs. Power (1971)"

Noam Chomsky, Michel Foucault

2006 | The Chomsky-Foucault Debate | DOI: 10.2307/jj.32047651.4

Źródło

[35] "The Power and Politics of Michel Foucault: An Interview with the Daily Californian"

Michel Foucault, Robert Joseph McCann

2026 | Foucault Studies | DOI: 10.1353/fou.00048

Źródło

[36] "Algorithms of Oppression: How Search Engines Reinforce Racism"

Safiya Noble

2018 | NYU Press eBooks | DOI: 10.2307/j.ctt1pwt9w5

Źródło

[37] "Social Justice as Topic and Tool: An Attempt to Transform an LIS Curriculum and Culture"

Nicole A. Cooke, Miriam E. Sweeney, Safiya Noble

2015 | The Library Quarterly | DOI: 10.1086/684147

Źródło

[38] "We're Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness"

Matthew Bui, Safiya Noble

2020 | Oxford University Press eBooks | DOI: 10.1093/oxfordhb/9780190067397.013.9

Źródło

[39] "Empowered to Name, Inspired to Act: Social Responsibility and Diversity as Calls to Action in the LIS Context"

Sarah Roberts, Safiya Noble

2016 | Library trends | DOI: 10.1353/lib.2016.0008

Źródło

[40] "Social Space and Symbolic Power"

Pierre Bourdıeu

1989 | Sociological Theory | DOI: 10.2307/202060

Źródło

[41] "Distinction: A Social Critique of the Judgement of Taste*"

Pierre Bourdıeu

2018 | Routledge eBooks | DOI: 10.4324/9781315680347-10

Źródło

[42] "Okonomisches Kapital, kulturelles Kapital, soziales Kapital. in Soziale Ungleichheiten. (Capital économique, culturel et social)"

Pierre Bourdıeu, Kreckel ed

1983

Źródło

[43] "The social space and the genesis of groups"

Pierre Bourdıeu

1985 | Theory and Society | DOI: 10.1007/bf00174048

Źródło

[44] "Autonomy, Interdependence, and Social Control: NASA and the Space Shuttle Challenger"

Diane Vaughan

1990 | Administrative Science Quarterly | DOI: 10.2307/2393390

Źródło

[45] "Rational Choice, Situated Action, and the Social Control of Organizations"

Diane Vaughan

1998 | Law & Society Review | DOI: 10.2307/827748

Źródło

[46] "Controlling Unlawful Organizational Behavior: Social Structure and Corporate Misconduct"

Diane Vaughan

1983 | DigitalGeorgetown (Georgetown University Library)

Źródło

[47] **"Organizational rituals of risk and error"**

Diane Vaughan

2005 | Cambridge University Press eBooks | DOI: 10.1017/cbo9780511488580.002

Źródło

Artykuły naukowe autora:

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

Chirag Shah

2026 | DOI: 10.5040/9798216452430

Źródło

[49] **"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: 17ef676c-78b7-459f-abb4-1299e8728581