Między kodem a sumieniem: test dojrzałości cyfrowej

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje ewolucję sztucznej inteligencji z prostego narzędzia w złożoną instytucję poznawczą i społeczną, która aktywnie kształtuje współczesny ład. Autor stawia tezę, że AI nie jest neutralna, lecz stanowi fundament nowego porządku, wymagający dojrzałego podejścia do etyki i odpowiedzialności. W tekście szczegółowo omówiono ramy regulacyjne, w tym unijny AI Act, standard ISO/IEC 42001 oraz wytyczne NIST AI RMF dotyczące zarządzania ryzykiem. Poruszono krytyczne zagadnienia, takie jak wyjaśnialność algorytmów (XAI), decision provenance oraz problematykę 'automatyzacyjnego lenistwa'. Tekst ostrzega przed ucieczką od odpowiedzialności w chmurę i wzywa do audytu algorytmicznego oraz walki z dyskryminacją. To kompleksowe spojrzenie na technologię jako element demokratycznego państwa prawa, gdzie skuteczność musi iść w parze ze sprawiedliwością.

This article examines the evolution of artificial intelligence from a simple tool into a complex cognitive and social institution that actively shapes the contemporary order. The author argues that AI is not neutral but rather constitutes the foundation of a new order, requiring a mature approach to ethics and responsibility. The text provides a detailed discussion of the regulatory framework, including the EU AI Act, the ISO/IEC 42001 standard, and the NIST AI RMF risk management guidelines. Critical issues such as explainability of algorithms (XAI), decision provenance, and the problem of "automation laziness" are addressed. The text warns against escaping responsibility into the cloud and calls for algorithmic auditing and the fight against discrimination. This is a comprehensive view of technology as an element of a democratic rule of law, where effectiveness must go hand in hand with justice.

Sztuczna inteligencja przestała być jedynie technologicznym narzędziem, stając się pełnoprawną instytucją kształtującą nasze życie społeczne, ekonomiczne i polityczne. Niniejszy artykuł podejmuje krytyczną analizę tej cyfrowej transformacji, odrzucając

naiwną metaforę AI jako neutralnego młotka. Autor stawia tezę, że wdrożenie systemów uczących się w sektorach krytycznych – takich jak medycyna, finanse czy administracja publiczna – wymaga fundamentalnej zmiany reżimu odpowiedzialności. Skuteczność algorytmiczna bez etycznego zakotwiczenia i mechanizmów audytowalności staje się jedynie przyspieszoną niesprawiedliwością. W obliczu zagrożeń płynących z asymetrii wiedzy, automatyzacyjnego lenistwa oraz erozji debaty publicznej przez syntetyczne treści, tekst postuluje przejście od deklaratywnej etyki do twardej architektury nadzoru. Kluczowym wyzwaniem staje się tutaj wiedza translacyjna, pozwalająca przekładać abstrakcyjne wartości na konkretne procedury techniczne i prawne. Ostatecznie artykuł dowodzi, że budowa zaufanej sztucznej inteligencji nie jest kwestią optymalizacji parametrów modelu, lecz kwestią ustrojową, w której innowacja musi zostać świadomie podporządkowana dobru wspólnemu, a nie jedynie logice bezrefleksyjnego wzrostu.

SŁOWA KLUCZOWE

sztuczna inteligencja | AI Act | etyka technologii | zarządzanie ryzykiem | wyjaśnialność AI | audyt algorytmiczny | NIST AI RMF | ISO/IEC 42001 | multimodalność | modele ogólnego przeznaczenia | automatyzacyjne lenistwo | decision provenance | dyskryminacja algorytmiczna | instytucja poznawcza | odpowiedzialność cyfrowa

Sztuczna inteligencja jako fundament nowego ładu społecznego

„Nie wystarczy być pracowitym; mrówki też są. Nad czym tak pracowicie się trudzisz?" To zdanie Henry'ego Davida Thoreau, przywołane przez Terezę Raquel Merlo, stanowi coś znacznie głębszego niż tylko elegancki ornament literacki na marginesie inżynieryjnego raportu. W istocie jest to pierwszy, fundamentalny test dojrzałości dla naszej cywilizacji algorytmicznej, która zdaje się pędzić przed siebie bez mapy. Możemy przecież bez końca trenować modele coraz potężniejsze, szybsze i bardziej multimodalne, czyli zdolne do jednoczesnego operowania na tekście, obrazie i dźwięku, co pozwala maszynom na niemal ludzką percepcję otoczenia. Osiągamy mistrzostwo w imitowaniu mowy, rozumowania czy estetyki, budując systemy obsługujące w sekundę więcej

komunikatów, niż wszystkie parlamenty świata wypowiedziały przez ostatnie stulecie. Niemniej jednak, gdy tracimy z oczu cel owego trudu, maszyna zamienia się w imponującą, lecz bezduszną kolonię mrówek, pracującą z nieludzką wydajnością nad procesami, których nikt już nie potrafi moralnie uzasadnić.

Główna teza, którą musimy tu z całą mocą wyartykułować, jest bezlitosna dla technosceptyków i optymistów zarazem: sztuczna inteligencja nigdy nie była i nie będzie neutralnym narzędziem. To nie jest po prostu kolejny młotek, który raz wbija gwóźdź w ścianę, a innym razem rozbija szybę, zależnie od chwilowego kaprysu czy intencji jego użytkownika. Taka uproszczona metafora jest nie tylko poznawczo uboga, ale wręcz niebezpiecznie dziecięca, choć w debacie publicznej wciąż pozostaje niezwykle wygodna, bo skutecznie zwalnia nas z obowiązku głębszego myślenia. AI jest bowiem pełnoprawną instytucją poznawczą, ekonomiczną, prawną oraz polityczną, która aktywnie współtworzy ramy naszej rzeczywistości. Skuteczność bez sprawiedliwości jest tylko przyspieszoną niesprawiedliwością, a my wciąż zbyt często udajemy, że patrzymy na zwykły, niegroźny śrubokręt.

Systemy uczące się nie ograniczają się do biernego wykonywania poleceń, lecz aktywnie klasyfikują świat, ustalają hierarchie prawdopodobieństwa i decydują o tym, co staje się widzialne. Rozdzielają one dostęp do kluczowych usług, filtrują strumienie informacji, a także z niepokojącą precyzją reprodukują lub – znacznie rzadziej – korygują nasze głęboko zakorzenione uprzedzenia. Kiedy algorytm zostaje osadzony w sektorze krytycznym, takim jak medycyna czy finanse, przestaje być zwykłą aplikacją, a staje się żywym fragmentem ładu społecznego. Właśnie dlatego publikacje o etyce technologii powinniśmy czytać jak traktaty o instytucjonalizacji odpowiedzialności, a nie jak instrukcje obsługi serwera. To tutaj rozstrzyga się, czy technologia pozostanie podległa dobru praktycznemu, czy też ostatecznie wymknie się nam z rąk pod płaszczem rzekomej optymalizacji.

Intuicja stojąca za odpowiedzialnym projektowaniem systemów jest w swej istocie arystotelesowska, choć dziś ubieramy ją w nowoczesny kostium nauki o danych. Technologia nabiera sensu dopiero wtedy, gdy zostaje świadomie podporządkowana celom wyższym, co wymaga przekształcenia abstrakcyjnych wartości w konkretne procedury, audyty i ścieżki decyzyjne. Etyka nie może być przecież kwiatkiem przypiętym do serwera, lecz musi stanowić integralny projekt konstrukcyjny całego mechanizmu społeczno-technicznego. Nie chodzi o ozdobne przymiotniki dopisane do dokumentacji technicznej, ale o realną architekturę danych i kulturę organizacyjną, która nie pozwala na ucieczkę od odpowiedzialności w chmurę. Właśnie w tym miejscu rodzi się współczesny paradygmat naukowy, który porzuca naiwną fascynację benchmarkami na rzecz dojrzałego myślenia o systemach złożonych.

W tym nowym ujęciu model statystyczny jest zaledwie jednym z wielu elementów skomplikowanej układanki, w której obok kodu istnieją infrastruktura, regulatorzy, rynki oraz historia

dyskryminacji. Musimy brać pod uwagę asymetrię wiedzy, ograniczenia poznawcze człowieka, a przede wszystkim wszechobecną pokusę automatyzacyjnego lenistwa, która każe nam ufać maszynie bardziej niż własnemu osądowi. Aluzja do starożytnego mitu o Golemie narzuca się tutaj niemal automatycznie, choć wymaga ona istotnej, współczesnej korekty. Dzisiejszy Golem nie jest już glinianym olbrzymem, któremu mędrzec wkłada pod język kartkę z magicznym słowem, lecz rozproszoną siecią modeli, interfejsów API oraz niewidzialnych zależności kontraktowych. To nie jest postęp; to jest ucieczka od odpowiedzialności w chmurę, której musimy się aktywnie przeciwstawić.

Od efektywności do odpowiedzialności: AI jako zmiana reżimu władzy

Największym zagrożeniem wcale nie jest mityczny moment, w którym krzemowa inteligencja nagle otworzy oczy i ogłosi swoją podmiotowość w dramatycznym geście buntu. Prawdziwe niebezpieczeństwo tkwi w tym, że ona już tu jest, działa pełną parą w samym sercu naszych instytucji, a my wciąż uparcie odgrywamy komedię, w której AI pozostaje jedynie neutralnym młotkiem w rękach rzemieślnika. Ta iluzja narzędziowości powoli jednak pęka, a najbardziej spektakularnym dowodem na tę zmianę jest dojrzewająca w Europie odpowiedź regulacyjna, która próbuje okiełznać cyfrowy żywioł. Unijny AI Act, czyli europejskie rozporządzenie o sztucznej inteligencji wprowadzające ramy prawne dla bezpiecznego rozwoju technologii, wszedł w życie 1 sierpnia 2024 roku i nie jest jedynie kolejnym nudnym dokumentem z Brukseli. Harmonogram jego wdrażania – od zakazu najbardziej inwazyjnych praktyk w lutym 2025 roku, przez rygory dla modeli ogólnego przeznaczenia w sierpniu tego samego roku, aż po pełne obowiązywanie zasad w 2026 i 2027 roku – to coś więcej niż biurokratyczny kalendarz. To sygnał, że era dobrowolnych deklaracji etycznych, które zbyt często przypominały kwiatek przypięty do serwera, bezpowrotnie dobiegła końca na rzecz twardego zarządzania ryzykiem.

Przejście od etyki do zarządzania ryzykiem publicznym oznacza, że AI Act traktuje algorytmy nie jako matematyczną ciekawostkę, lecz jako potencjalne źródło systemowych wstrząsów o charakterze politycznym. Logika tego aktu opiera się na trzeźwym założeniu, że nie każda linijka kodu jest równie groźna, niemniej niektóre z nich bezpośrednio dotykają fundamentów naszego życia: zdrowia, bezpieczeństwa, a nawet samej demokracji. Kiedy systemy decydują o naszym zatrudnieniu, dostępie do edukacji, infrastrukturze krytycznej czy wyrokach sądowych, przestają być problemem czysto technicznym, a stają się kwestią ustrojową. Oficjalne ujęcie aktu stawia sobie za cel wspieranie rozwoju sztucznej inteligencji zorientowanej na człowieka, co w języku

prawniczym oznacza próbę ochrony praw podstawowych przed bezduszną, zautomatyzowaną optymalizacją. Jest to kluczowy moment, w którym język prawa zaczyna nazywać to, co praktycy danych często woleli spychać na margines: system AI nie jest wyłącznie wyzwaniem dla efektywności, lecz nowym polem bitwy o kształt naszych instytucji.

W tym nowym porządku audyt algorytmiczny przestaje być nudnym, rytualnym odhaczaniem listy kontrolnej tuż przed premierą produktu, a staje się odpowiednikiem konstytucyjnego mechanizmu kontroli władzy. W dojrzałej demokracji nie ufamy władzy wykonawczej tylko dlatego, że obiecała działać rozsądnie, toteż w świecie technologii nie możemy polegać wyłącznie na zapewnieniach dostawców, że ich model jest „fair", „safe" czy „responsible". Stara maksyma mówi, że władza bez kontroli korumpuje ludzi, jednak w dobie algorytmów musimy zrozumieć, że władza bez audytu korumpuje same modele, procesy i całe struktury organizacyjne. Owa korupcja rzadko bywa efektem złej woli programisty, częściej przybierając formę podstępnego dryfu danych, źle zdefiniowanej funkcji celu czy automatyzacyjnego uprzedzenia użytkownika, który bezkrytycznie ufa maszynie. Może to być również nieszczelność pamięci lub elegancki dashboard, który estetycznie maskuje narastającą katastrofę epistemiczną, czyli sytuację, w której tracimy zdolność do odróżnienia rzetelnej wiedzy od błędnych założeń systemu.

Aby uniknąć tego rodzaju pułapek, NIST AI Risk Management Framework porządkuje ten sposób myślenia poprzez cztery funkcje: Govern, Map, Measure i Manage, tworząc praktyczną gramatykę odpowiedzialności. Funkcja Govern wymaga od organizacji stworzenia trwałych struktur, ról i kultury zarządzania ryzykiem, zamiast polegania na doraźnej improwizacji w sytuacjach kryzysowych. Map nakazuje precyzyjne rozpoznanie kontekstu, zastosowania oraz potencjalnych szkód dla wszystkich interesariuszy, podczas gdy Measure wymusza rygorystyczne testowanie i monitorowanie wyników w oparciu o obiektywne kryteria. Ostatni element, Manage, dotyczy konkretnych decyzji o ograniczaniu, przenoszeniu lub dokumentowaniu ryzyka, co w praktyce przesuwa ciężar dyskusji z pytania o to, czy model działa, na pytanie o to, co to właściwie oznacza dla człowieka. Oficjalny NIST podkreśla wszak, że te ramy służą lepszemu zarządzaniu ryzykiem wobec jednostek i społeczeństwa, a nie tylko szlifowaniu technicznych parametrów, które często bywają jedynie zasłoną dymną dla braku odpowiedzialności.

Musimy wreszcie zrozumieć, że model rekrutacyjny może być statystycznie bezbłędny, a jednocześnie w sposób bezlitosny utrwalać wykluczenia klasowe, płciowe czy kulturowe, których nie widać w prostym zestawieniu wyników. Podobnie scoring algorytmiczny, czyli zautomatyzowana ocena wiarygodności stosowana w bankowości, może poprawiać trafność predykcji, a jednocześnie zamykać drogę do kapitału grupom historycznie niedoreprezentowanym. System medyczny może imponować wykrywalnością chorób, a jednocześnie systematycznie gorzej

obsługiwać populacje, które z racji zaszłości były słabiej reprezentowane w zbiorach treningowych, co prowadzi do ukrytej dyskryminacji. Nawet chatbot wyborczy, odpowiadający z nienaganną uprzejmością w wielu językach, może stać się narzędziem chaosu, jeśli w kluczowym momencie poda błędną informację o miejscu głosowania lub terminie wyborów. To właśnie w takich momentach umiera naiwna definicja skuteczności, bo skuteczność bez sprawiedliwości jest przecież tylko przyspieszoną niesprawiedliwością, ubraną w szaty nowoczesnej technologii.

Odpowiedzią na te wyzwania z perspektywy organizacyjnej jest standard ISO/IEC 42001:2023, który zamiast na jednorazowych zrywach, skupia się na budowaniu zintegrowanego systemu zarządzania sztuczną inteligencją. Standard ten ma pomagać organizacjom w odpowiedzialnym używaniu AI poprzez ocenę ryzyk w całym cyklu życia systemu, co przypomina logikę dojrzałych systemów zarządzania bezpieczeństwem informacji czy prywatnością. Dla biznesu jest to punkt zwrotny, ponieważ wiele organizacji wciąż postrzega AI jako efektowny dodatek produktywnościowy, mający jedynie zredukować koszty i ładnie wyglądać na slajdach podczas prezentacji dla zarządu. Tymczasem w sektorach krytycznych wdrożenie algorytmu nie jest zakupem nowego narzędzia, lecz fundamentalną zmianą reżimu odpowiedzialności, która rekonfiguruje relację władzy między instytucją a obywatelem. Jeżeli system współdecyduje o kredycie, diagnozie medycznej czy selekcji kandydata do pracy, to organizacja musi zapewnić kontrolę, jawność oraz realną możliwość wyjaśnienia i odwołania się od decyzji. Prometeusz nie musi oddawać ognia, ale musi wreszcie przestać udawać, że ten ogień nie parzy i że nie ponosi odpowiedzialności za pożary, które może wzniecić w naszym wspólnym domu.

Algorytmiczna administracja: między audytem a odpowiedzialnością

Bez przejrzystości nie mamy do czynienia z innowacją, lecz z formą administracyjnej magii, która wymyka się racjonalnemu osądowi i demokratycznej kontroli. Znany bon mot przypomina nam dobitnie, że model sztucznej inteligencji, którego nie można wyjaśnić, to po prostu model, któremu nie wolno ufać. Można tę myśl jeszcze bardziej zaostrzyć: algorytmu, którego działania nie potrafimy zrekonstruować w sprawach dotyczących praw człowieka, nie wolno używać tak, jakby był ostatecznym wyrokiem. Wyjaśnialność, czyli XAI, będąca zdolnością systemu do przedstawienia logicznych przesłanek decyzji, nie jest jedynie epistemologicznym luksusem dla znudzonych akademików. Stanowi ona fundamentalny warunek jakiejkolwiek kontroli nad procesami, które kształtują naszą rzeczywistość społeczną i prawną. Innowacja bez zaufania jest tylko pokazem pirotechnicznym w magazynie benzyny.

Jeżeli konkretna decyzja administracyjna nie może zostać w pełni zrekonstruowana, to w konsekwencji nie może zostać skutecznie zakwestionowana przez obywatela. W takim scenariuszu algorytm przestaje być narzędziem pomocniczym, a staje się formą czystej przemocy proceduralnej, ukrytej pod płaszczem nowoczesnej technologii. Gdy człowiek słyszy jedynie beznamiętny komunikat, że „system tak wskazał", przestaje uczestniczyć w racjonalnej administracji państwowej. Zamiast tego staje się mimowolnym świadkiem technokratycznego misterium, gdzie tradycyjny ołtarz zastąpiono lśniącym interfejsem, a dawną wyrocznię nowoczesnym pulpitem analitycznym. To właśnie w tym punkcie etyka sztucznej inteligencji spotyka się z twardym prawem administracyjnym i klasyczną teorią państwa.

Demokratyczna instytucja, w przeciwieństwie do prywatnego startupu, nie może ograniczać się jedynie do statystycznej trafności uzyskanego wyniku. Musi ona posiadać zdolność do merytorycznego uzasadnienia każdej podjętej decyzji, która wpływa na życie jednostki. Musi zapewnić proporcjonalność, równe traktowanie, realną możliwość odwołania oraz ścisłą kontrolę proceduralną na każdym etapie procesu. Algorytmiczna administracja pozbawiona tych gwarancji staje się państwem administracyjnym w wersji cyfrowej: szybkim, elastycznym i pozornie racjonalnym, lecz potencjalnie nieprzenikalnym dla zwykłego obywatela. W tym sensie technologia, zamiast wyzwalać, może stać się nowym narzędziem wykluczenia, które operuje poza zasięgiem wzroku opinii publicznej.

Dawna biurokracja, mimo wszystkich swoich wad i ociężałości, przynajmniej zostawiała po sobie wyraźny, fizyczny ślad papierowy. Nowa biurokracja może natomiast zostawić ślad jedynie w logach, których nikt poza dostawcą technologii nie rozumie, albo których w ogóle nie zaprojektowano z myślą o audycie. Takie rozwiązanie nie jest postępem, lecz ucieczką od odpowiedzialności w chmurę, gdzie winowajcę trudniej wskazać niż w labiryncie korytarzy Kafki. Dlatego tak kluczowe stają się mechanizmy decision provenance, czyli systematycznego śledzenia pochodzenia każdej decyzji. W systemach agentowych, korzystających z pamięci i autonomicznego planowania, musimy dokładnie wiedzieć, skąd dana konkluzja się wzięła i jakie przesłanki o niej przesądziły.

Bez precyzyjnej genealogii decyzji każda organizacja posiada jedynie performans kontroli, będący pustym rytuałem bez realnego znaczenia. Musimy wiedzieć, które konkretnie dane zostały użyte, który model wygenerował rekomendację i jakie polityki bezpieczeństwa zostały w danym momencie aktywowane. Należy monitorować, jakie wywołania narzędzi nastąpiły, czy pamięć systemu została zmodyfikowana oraz czy człowiek ostatecznie zatwierdził uzyskany wynik. Trzeba rozstrzygnąć, czy rezultat był efektem predykcji, sztywnej reguły, heurystyki, czy może zwykłej halucynacji lub błędu

integracji. W rzeczywistości organizacja bez takiej wiedzy stoi przed własnym systemem jak uczeń przed mitycznym Sfinksem: słyszy odpowiedź, ale nie zna drogi rozumowania.

Standard IEEE CertifAIEd wpisuje się w tę samą architekturę dojrzałości, oferując program oceny etyki autonomicznych systemów inteligentnych. Ma on wspierać budowanie bardziej godnego zaufania doświadczenia użytkowników i wykazywać, że dany produkt został rzetelnie oceniony pod kątem zasad etycznych. Tego rodzaju certyfikacja oczywiście nie rozwiązuje wszystkich problemów, bo żadna pieczęć nie zastąpi realnej kultury odpowiedzialności wewnątrz firmy. Tworzy jednak niezwykle ważny język wspólnej oceny, który pozwala na dialog między inżynierami a humanistami. Organizacje potrzebują dziś nie tylko ogólnych wartości, lecz także ontologii wartości, czyli wspólnych pojęć, kryteriów i akceptowalnych progów ryzyka.

Ryzyko algorytmiczne należy przy tym rozumieć znacznie szerzej niż tylko jako prosty błąd techniczny czy awarię serwera. Błąd techniczny jest bowiem tylko jednym z wielu gatunków ryzyka, które mogą zainfekować systemy oparte na danych. Stronniczość algorytmiczna może wynikać z danych historycznych, które w istocie są jedynie cyfrowym archiwum dawnych niesprawiedliwości i uprzedzeń. Z kolei reward misalignment, czyli niedopasowanie nagrody, może sprawić, że agent będzie optymalizował suchy wskaźnik, całkowicie ignorując nadrzędny cel społeczny. To klasyczny problem, w którym maszyna wykonuje polecenie dosłownie, ale w sposób sprzeczny z intencją jej twórcy.

W systemach tych może pojawić się również deceptive reasoning, czyli zwodnicze wnioskowanie, gdzie system uczy się osiągać rezultat przez sprytne obchodzenie intencji nadzorcy. Innym zagrożeniem jest automation bias, powodujący, że człowiek, zamiast sprawować realny nadzór, jedynie ceremonialnie zatwierdza każdą rekomendację maszyny. Nie możemy też zapominać o dryfie modelu, który sprawia, że system skuteczny w dniu wdrożenia staje się niebezpieczny po zmianie populacji lub środowiska prawnego. Ktoś mógłby powiedzieć, że to tylko ryzyka operacyjne, ale w rzeczywistości jest to nowa anatomia władzy w nowoczesnej organizacji. Skuteczność bez sprawiedliwości jest tylko przyspieszoną niesprawiedliwością.

Audyt systemów AI musi być procesem ciągłym, a nie jednorazowym wydarzeniem odnotowanym w kalendarzu. Jednorazowy test przed wdrożeniem jest w świecie sztucznej inteligencji tym, czym jednorazowe badanie lekarskie w dzieciństwie dla pilota samolotu pasażerskiego po czterdziestu latach służby. Może mieć ono pewną wartość historyczną, ale w żaden sposób nie gwarantuje bieżącego bezpieczeństwa pasażerów i załogi. Modele funkcjonują w środowisku skrajnie zmiennym, gdzie dane ulegają nieustannemu przesunięciu, a użytkownicy błyskawicznie uczą się obchodzić systemowe zabezpieczenia. Adwersarze atakują promptami, zatruwają dane treningowe i testują granice wytrzymałości infrastruktury w sposób niezwykle kreatywny.

Audyt statyczny jest więc jedynie konserwą odpowiedzialności, która szybko traci termin przydatności do spożycia. Audyt dojrzały musi być procesem żywym, reagującym na bieżąco na zmieniające się warunki zewnętrzne i wewnętrzne. Stąd bierze się ogromne znaczenie red teamingu, czyli kontrolowanych ataków mających na celu wykrycie słabych punktów systemu. Testy kontradyktoryjne nie są przejawem braku zaufania wobec technologii, ale formą głębokiego szacunku wobec jej ogromnej mocy. System, którego nikt nie próbuje złamać przed wdrożeniem, zostanie z pewnością złamany po wdrożeniu przez kogoś znacznie mniej uprzejmego niż audytor.

Red teaming pozwala sprawdzić podatność na prompt injection, manipulację kontekstem, wycieki danych oraz generowanie treści szkodliwych. W sektorach krytycznych, takich jak ochrona zdrowia czy finanse, powinien być on tak naturalny jak próby obciążeniowe nowo wybudowanego mostu. Nikt rozsądny nie mówi przecież: „most wygląda ładnie, więc nie psujmy radosnej atmosfery brutalnymi testami wytrzymałościowymi". W świecie AI wciąż jednak zbyt często słyszymy tę specyficzną wersję optymizmu, ubraną w modny polar startupowy i żargon innowacji. Tymczasem audyt bez odpowiedniej infrastruktury jest jedynie pustą deklaracją, która nie ma pokrycia w rzeczywistych działaniach.

Współczesna debata publiczna kocha mówić o jednostkach GPU, ponieważ są one efektowne, drogie i brzmią jak potężne konie pociągowe nowej epoki. Jednak zaufana sztuczna inteligencja wymaga przede wszystkim zaufanych danych, a te wymagają infrastruktury pozwalającej ustalić ich pochodzenie, jakość i integralność. Należy precyzyjnie określić okres przechowywania danych, zgodę na ich użycie oraz podstawę prawną przetwarzania w konkretnym kontekście. Hasło, że zaufana sztuczna inteligencja zaczyna się od zaufanych danych, nie jest jedynie chwytem marketingowym. To fundamentalna zasada ustrojowa nowoczesnej gospodarki algorytmicznej, której nie wolno nam ignorować.

Problem polega na tym, że dane w systemach AI nie są paliwem, które spala się jednorazowo w komorze silnika. Tworzą one raczej skomplikowaną pętlę: surowe dane zasilają model, model generuje wyniki, a te wyniki stają się nowymi danymi dla kolejnych procesów. To właśnie nieskończona pętla danych AI, w której błędne założenie może błyskawicznie stać się samospełniającą się przepowiednią. Jeżeli system kredytowy historycznie gorzej ocenia pewną grupę, algorytm zaczyna przypominać mitologicznego Kronosa, który pożera własne dzieci. Następnie system ten twierdzi, że przecież statystyka głodu jest całkowicie obiektywna, ignorując fakt, że sam ją aktywnie współtworzy.

Dlatego zarządzanie wiedzą, czyli Knowledge Management, nie jest jedynie opcjonalnym dodatkiem do sztucznej inteligencji, lecz jej systemem nerwowym. Musimy pamiętać, że same dane nie są jeszcze wiedzą, model nie jest jeszcze rozumieniem, a predykcja nie jest jeszcze ostatecznym

osądem. Organizacja, która nie potrafi rozróżnić tych pojęć, buduje jedynie bardzo drogie urządzenie do wzmacniania własnych złudzeń i błędów. Zarządzanie wiedzą pozwala uchwycić wiedzę jawną i ukrytą, a następnie przekształcać ją w sposób użyteczny dla systemów maszynowych. To tutaj pojawia się kluczowe pojęcie wiedzy translacyjnej, czyli zdolności przekładania ludzkich wartości na parametry techniczne.

Bez wiedzy translacyjnej etyka pozostaje jedynie pięknym kazaniem wygłoszonym do kompilatora – wzruszającym, ale technicznie całkowicie bezskutecznym. Z wiedzą translacyjną staje się ona natomiast twardym wymaganiem projektowym, które musi zostać spełnione przez inżynierów. Jeżeli mówimy, że system rekrutacyjny ma być sprawiedliwy, musimy określić, jakie dane wykluczamy i jakie testy fairness stosujemy. Należy zdefiniować, jakie progi odchyleń są akceptowalne oraz jakie konkretne decyzje wymagają bezwzględnego udziału człowieka. Etyka, która nie przechodzi przez taki proces przekładu na język techniczny, jest tylko dobrym wychowaniem prezentowanym na branżowych konferencjach.

Podobnie sytuacja wygląda w systemach medycznych, gdzie bezpieczeństwo musi zostać zdefiniowane poprzez konkretne procedury i ograniczenia. Trzeba jasno określić, kiedy rekomendacja AI jest tylko sugestią, a kiedy musi zostać zignorowana przez lekarza prowadzącego. Należy ustalić, jak mierzymy błąd dla różnych populacji i kto ostatecznie bierze na siebie odpowiedzialność za decyzję kliniczną. Etyka nie może być kwiatkiem przypiętym do serwera, lecz musi stanowić integralną część architektury systemu od samego początku jego projektowania. Tylko wtedy możemy mówić o technologii, która rzeczywiście służy człowiekowi, a nie jedynie optymalizuje zyski korporacji.

Refleksja ta prowadzi nas do koncepcji human-in-the-loop, która nie może być rozumiana jedynie jako wygodne alibi dla twórców algorytmów. Człowiek „w pętli” bywa czasem tylko zmęczonym operatorem przy guziku, który pod presją czasu i wskaźników KPI zatwierdza każdą decyzję maszyny. Prawdziwy nadzór ludzki wymaga kompetencji, czasu, dostępu do pełnego uzasadnienia oraz kultury organizacyjnej, która nie karze za sceptycyzm. Inaczej człowiek w pętli jest jak strażnik w muzeum, któremu zgaszono światło i zabrano okulary, każąc pilnować Rembrandta przez kamerę z opóźnieniem. Obywatel nie traci wolności w dramatycznej scenie z muzyką Wagnera; traci ją po cichu, scrollując i akceptując nieprzejrzyste decyzje systemów, których nie rozumie. Ta refleksja nad odpowiedzialnością i kontrolą prowadzi nas nieuchronnie w stronę pytań o przyszłość samej demokracji.

Demokracja w epoce algorytmów: między innowacją a odpowiedzialnością

Rok 2024 zapisał się w kronikach jako globalny poligon doświadczalny, na którym sprawdzano odporność sfery publicznej na uderzeniową falę generatywnej sztucznej inteligencji. Według raportu Reuters Institute z marca 2024 roku w blisko pięćdziesięciu procesach wyborczych udział mogło wziąć około dwóch miliardów ludzi, co sprowokowało fundamentalne pytanie o to, czy syntetyczna dezinformacja jest w stanie realnie wykrzywić kręgosłup demokracji. Nie chodziło tu jednak o spektakularne przewroty wywołane pojedynczym cyfrowym kłamstwem, lecz o systematyczną erozję tego, co nazywamy minimum wspólnej rzeczywistości. Gdy przestrzeń debaty zostaje zalana potopem syntetycznych obrazów, spreparowanych głosów i botów udających autorytety, obywatel rzadko kiedy zaczyna bezkrytycznie wierzyć w każdą bzdurę. Znacznie częściej popada on w paraliżujący marazm, przestając wierzyć, że odróżnienie prawdy od fałszu jest w ogóle możliwe, co stanowi dla ustroju truciznę znacznie subtelniejszą i groźniejszą niż prymitywna propaganda. Niemniej, zachowując intelektualną uczciwość, nie możemy ulegać apokaliptycznemu automatyzmowi, zwłaszcza gdy giganci tacy jak Meta sugerują w swoich podsumowaniach, że wpływ dezinformacji na ich platformach był w 2024 roku raczej skromny. Tę korektę należy oczywiście czytać z dużą dozą ostrożności, wszak pochodzi ona od podmiotu bezpośrednio zainteresowanego tym, jak oceniane jest ryzyko jego własnych usług. Ostateczny wniosek nie brzmi więc: „AI zniszczyła wybory", lecz jest znacznie bardziej niepokojący: technologia ta drastycznie obniżyła koszt produkcji niepewności, a nasze instytucje wciąż mozolnie uczą się, jak tę mgłę mierzyć i rozpraszać.

Choć deepfake, czyli syntetyczny materiał audiowizualny generowany przez AI w celu realistycznego imitowania rzeczywistości, przyciąga uwagę swoją spektakularnością, to wcale nie on musi być najbardziej niszczycielskim narzędziem w arsenale manipulatorów. Równie destrukcyjną rolę odgrywają tak zwane cheapfakes, masowa produkcja komentarzy, klonowanie autorytetów czy syntetyczny astroturfing, polegający na tworzeniu sztucznych, oddolnych ruchów społecznych, które mają sprawiać wrażenie autentycznego poparcia. Demokracja nie kruszy się pod wpływem jednego potężnego uderzenia, lecz zostaje rozbita przez milion małych zadrapań poznawczych, wynikających z mikrotargetowanych komunikatów, których nikt poza wybraną grupą odbiorców nie widzi. Obywatel nie traci swojej wolności w wielkiej, dramatycznej scenie z muzyką Wagnera w tle; on traci ją niemal niezauważalnie, bezrefleksyjnie scrollując ekran swojego smartfona. To nowoczesna, cyfrowa wersja jaskini Platona, w której cienie na ścianie nie są już rzucane przez migoczący ogień, lecz renderowane przez zaawansowane modele, optymalizowane pod kątem zaangażowania i bezlitośnie rozliczane w metrykach naszej uwagi.

Warto jednak pamiętać, że generatywna sztuczna inteligencja nie jest wyłącznie demonem czyhającym na nasze głosy, gdyż posiada ona ogromny potencjał wspierania inkluzywności i wielojęzyczności w debacie publicznej. Może ona służyć jako potężne narzędzie tłumaczenia treści wyborczych, komunikacji z mniejszościami narodowymi czy analizy realnych potrzeb obywateli, czyniąc państwo bardziej słyszalnym i – co ważniejsze – bardziej słuchającym. Personalizacja treści, choć bywa kuzynką propagandy, jest jednocześnie siostrą inkluzywności, pozwalając dotrzeć z informacją publiczną do osób z niepełnosprawnościami czy obywateli oddalonych od centrów decyzyjnych. Różnica między tymi dwoma obliczami technologii nie tkwi w samym kodzie, lecz w przejrzystości celu, zgodzie użytkownika oraz mechanizmach kontroli i odpowiedzialności. Dlatego właśnie fundamentalne pytanie zmieniło się z technicznego „czy system może zostać zbudowany" na etyczne „czy powinien zostać zbudowany", co powinno stać się mottem każdego laboratorium AI i każdego ministerstwa cyfryzacji.

Sama możliwość techniczna nie stanowi jeszcze legitymacji moralnej, a współczesny świat musi zmierzyć się z kumulacją czterech potężnych pokus: ekonomicznej maksymalizacji, prawnego formalizmu, politycznej kontroli oraz technologicznego pędu do wdrożenia. System może być przecież wysoce opłacalny, formalnie zgodny z minimalnymi wymogami prawa, użyteczny dla władzy i technicznie imponujący, a mimo to pozostawać głęboko szkodliwy dla tkanki społecznej. To właśnie z tego powodu etyka nie może być jedynie kwiatkiem przypiętym do serwera, lecz musi zostać wpisana w samą architekturę systemów, a audyt stać się integralną częścią ich cyklu życia. Tu kończy się beztroskie dzieciństwo sztucznej inteligencji, w którym pytaliśmy głównie o to, co algorytmy potrafią, a zaczyna się wiek dojrzały, wymuszający pytania o to, komu służą, komu zagrażają i kto ponosi realną odpowiedzialność za ich błędy.

Odpowiedzialna sztuczna inteligencja nie jest ani przejawem technofobii, ani bezkrytycznej technofilii, lecz stanowi stanowczą odmowę kapitulacji ludzkiego rozumu przed własnym, potężnym wynalazkiem. Frank Herbert ostrzegał niegdyś, że powierzając myślenie maszynom w nadziei na wyzwolenie, pozwalamy jedynie innym ludziom z maszynami na nasze zniewolenie, co w dobie modeli generatywnych przestaje być literacką dystopią, a staje się instrukcją bezpieczeństwa. Jest to szczególnie istotne w sektorach krytycznych, czyli miejscach, gdzie błąd algorytmu przestaje być jedynie statystyczną anomalią, a staje się nieodwołalnym losem konkretnego człowieka. W finansach, medycynie, cyberbezpieczeństwie czy administracji publicznej algorytm nie jest już tylko cyfrowym asystentem, lecz aktywnym uczestnikiem procesu rozdziału szans, praw, zasobów i bezpieczeństwa.

Kto w tych dziedzinach mówi wyłącznie o automatyzacji, ten w istocie głosi niebezpieczną półprawdę, która bywa groźniejsza niż jawny błąd, ponieważ nosi nienaganny garnitur

profesjonalizmu. Różnica między błędem AI w sektorze konsumenckim a pomyłką w sektorze krytycznym jest fundamentalna: w pierwszym przypadku użytkownik najwyżej straci wieczór na oglądaniu źle dobranego filmu, w drugim zaś może stracić zdrowie lub wolność. Jeżeli system medyczny, wykorzystujący cyfrowego bliźniaka, czyli wirtualny model pacjenta pozwalający na symulowanie skutków leczenia, błędnie oszacuje ryzyko sepsy, stawką staje się ludzkie życie. Podobnie dzieje się, gdy scoring algorytmiczny, czyli zautomatyzowana ocena wiarygodności stosowana w bankowości, ukryje dyskryminację za parawanem matematycznej neutralności, odcinając rodzinę od możliwości posiadania własnego mieszkania.

Skuteczność bez sprawiedliwości jest tylko przyspieszoną niesprawiedliwością, dlatego wdrożenie AI w obszarach wysokiego ryzyka wymaga stworzenia rygorystycznej architektury nadzoru i pełnej transparentności. Systemy te muszą być nie tylko sprawne, ale przede wszystkim audytowalne i wyjaśnialne, co w ramach XAI, czyli zdolności algorytmu do przedstawienia logicznych przesłanek swoich decyzji, umożliwia realną kontrolę obywatelską. Musimy pamiętać, że innowacja bez zaufania jest jedynie pokazem pirotechnicznym w magazynie benzyny, a technologia pozbawiona procedur odwoławczych staje się formą cyfrowego despotyzmu. „Jeśli nie potrafisz tego wyjaśnić, nie możesz tego uzasadnić” – to zdanie nie powinno być traktowane jako elegancki aforyzm na konferencyjne panele, lecz jako twardy wymóg prawny i moralny.

Wprowadzenie AI Act, czyli europejskiego rozporządzenia o sztucznej inteligencji wyznaczającego globalne standardy bezpieczeństwa, jest krokiem w stronę ucywilizowania tej cyfrowej dżungli, ale same przepisy nie zastąpią czujności. Musimy stale pytać, kto zyskuje przewagę dzięki algorytmom, kto zostaje uciszony przez ich działanie i kto ostatecznie płaci za błędy, których nikt nie chciał przewidzieć. Prometeusz nie musi oddawać ognia, ale musi wreszcie przestać udawać, że ogień nie parzy, zwłaszcza gdy płomień ten dotyka fundamentów naszego życia społecznego i politycznego. Ostatecznie to nie technologia nas zbawi lub potępi, lecz nasza zdolność do narzucenia jej ram odpowiedzialności, które sprawią, że innowacja będzie służyć człowiekowi, a nie tylko optymalizacji jego wykluczenia.

Finanse pod presją algorytmów: między innowacją a ryzykiem

To zasada rządów prawa, którą niemal niepostrzeżenie przetłumaczono na binarny język modeli predykcyjnych, czyniąc z finansów poligon doświadczalny dla cyfrowej rewolucji. Branża ta od dekad żyje z precyzyjnego pomiaru ryzyka, wypracowując hermetyczny żargon scoringu, ratingu czy Value at Risk, czyli metody szacowania maksymalnej straty w określonym czasie. Bankowość i

ubezpieczenia, uzbrojone w procedury AML i KYC, wydają się środowiskiem naturalnie skrojonym pod analitykę predykcyjną, wszak algorytmy czują się tu jak w domu. Niemniej właśnie w tym sterylnym świecie liczb ujawnia się fundamentalny paradoks: im bardziej sektor jest zaawansowany w kwantyfikacji, tym łatwiej myli on mierzalność z głębokim rozumieniem zjawisk. Innowacja bez zaufania jest przecież tylko pokazem pirotechnicznym w magazynie benzyny, a finanse to magazyn wyjątkowo łatwopalny.

Model może z chirurgiczną precyzją wyłapywać anomalie transakcyjne, lecz jednocześnie milczeć, gdy zapytamy go o logiczne uzasadnienie decyzji, co uderza w fundament wyjaśnialności. Wyjaśnialność (XAI), czyli zdolność systemu do przedstawienia przesłanek swojego werdyktu, staje się warunkiem koniecznym, by algorytmiczny wyrok nie był arbitralny. Algorytmy trafnie przewidują niewypłacalność, lecz często czynią to, korzystając z cech zastępczych, które po cichu reprodukują dawne nierówności klasowe, terytorialne czy pokoleniowe. W tym krajobrazie AI obiecuje nam szybsze wykrywanie nadużyć i tańszą obsługę klienta, przy czym każda z tych obietnic rzuca długi, niepokojący cień. Szybkość może bowiem zrodzić automatyczną podejrzliwość wobec grup słabiej reprezentowanych w danych, a tańsza obsługa oznacza, że człowiek w kryzysie rozmawia z systemem o empatii syntaktycznej. Chatbot bankowy przeprosi nas tonem aksamitnym i kojącym, lecz nie poniesie żadnej odpowiedzialności moralnej za decyzję, którą właśnie ukrył za parawanem procedury.

Aktualne obawy nadzorcze wykraczają jednak daleko poza kwestię uprzejmych algorytmów, ponieważ cyberzagrożenia ewoluują w tempie, któremu tradycyjne struktury kontrolne nie potrafią sprostać. Europejski nadzór rynków finansowych alarmuje, że AI drastycznie zwiększa złożoność ataków, co wymusza radykalne wzmocnienie kontroli nad instytucjami i ich technologicznymi dostawcami. Raporty z 2026 roku sygnalizują przy tym pogłębiającą się lukę kompetencyjną, wszak regulatorzy często nie zbierają jeszcze systematycznych danych o użyciu AI, podczas gdy banki wdrażają modele z prędkością światła. Oznacza to, że rynek dysponuje już potężną bronią poznawczą, której strażnik nie potrafi jeszcze ani nazwać, ani tym bardziej skutecznie rozbroić w razie potrzeby. Nadzór pozbawiony realnej kompetencji technicznej staje się jedynie kosztownym teatrem kontroli, a w finansach kurtyna opada zazwyczaj dopiero wtedy, gdy wybucha kryzys.

Nie mamy tu do czynienia z odruchem antyrynkowym, lecz z postulatem elementarnej symetrii władzy, bez której system finansowy traci swoją stabilność i wiarygodność. Jeżeli instytucje używają sztucznej inteligencji do optymalizacji portfeli i oceny klientów, to regulator musi rozumieć te procesy na poziomie operacyjnym, a nie tylko teoretycznym. Historia finansów wielokrotnie nas uczyła, że innowacja pozbawiona przejrzystości potrafi produkować nie tyle powszechne bogactwo, ile elegancko opakowane ryzyko systemowe. AI może stać się kolejnym instrumentem tej samej

bolesnej lekcji, jeśli audyt nie nadąży za postępującą automatyzacją, pozostawiając nas z iluzją bezpieczeństwa. Skuteczność bez sprawiedliwości jest przecież tylko przyspieszoną niesprawiedliwością, co w świecie cyfrowych finansów nabiera wyjątkowo dosłownego i groźnego znaczenia.

Wyjątkowo niebezpiecznym zjawiskiem jest postępująca koncentracja dostawców, gdzie dziesiątki banków i ubezpieczycieli opierają swoje fundamenty na tych samych modelach i infrastrukturze chmurowej. W takim układzie ryzyko przestaje być rozproszone i staje się zsynchronizowane, co przypomina sytuację, w której wszyscy pasażerowie statku nagle biegną na tę samą burtę. Dawniej obawiano się, że inwestorzy używający podobnych strategii mogą w chwili stresu rynkowego jednocześnie wyprzedawać aktywa, dziś zaś ta synchronizacja dotyczy cyberobrony i scoringu. Scoring algorytmiczny, czyli zautomatyzowana ocena wiarygodności stosowana powszechnie w bankowości, sprawia, że błąd jednego modelu może odciąć od kapitału całe grupy społeczne. Gdzie wszystkie modele myślą tak samo, tam nikt nie myśli zbyt wiele, a błąd jednostkowy błyskawicznie przepoczwarza się w paraliżujący kryzys całego systemu. Musimy zatem pamiętać, że technologiczny ogień, choć daje nam potężne światło, potrafi dotkliwie parzyć tych, którzy niosą go bez odpowiednich zabezpieczeń.

Od medycyny po przemysł: granice odpowiedzialności algorytmów

Opieka zdrowotna stanowi obszar wyjątkowo wymagający dla technologii, ponieważ w sposób bezpośredni dotyka ona materii ludzkiego ciała, doświadczenia bólu, nieuchronności śmierci oraz fundamentu, jakim jest zaufanie. W tym kontekście sztuczna inteligencja może okazać się jednym z najpotężniejszych sprzymierzeńców gatunku ludzkiego, wspierając diagnostykę obrazową, analizując ogrom dokumentacji czy precyzyjnie pomagając w procesach triażu. Systemy te potrafią monitorować pacjentów zmagających się z chorobami przewlekłymi, przewidywać ryzyko nagłego pogorszenia stanu zdrowia, a także radykalnie przyspieszać badania nad nowymi lekami i optymalizować skomplikowaną logistykę szpitalną. Cyfrowy bliźniak, czyli wirtualny model pacjenta lub organu pozwalający na symulowanie przebiegu chorób i skutków leczenia, w domowej opiece zdrowotnej może nieustannie nadzorować parametry życiowe i koordynować przyjmowanie leków. W obliczu wyzwań starzejących się społeczeństw takie rozwiązanie przestaje być technologicznym luksusem, stając się warunkiem koniecznym dla zachowania powszechnej dostępności opieki medycznej.

W świecie medycyny każde wprowadzone narzędzie obciążone jest ciężarem przysięgi Hipokratesa, co sprawia, że innowacja bez zaufania staje się jedynie pokazem pirotechnicznym w magazynie benzyny. Światowa Organizacja Zdrowia, reagując na te wyzwania, opublikowała w styczniu 2024 roku kompleksowe wytyczne dotyczące etyki i zarządzania dużymi modelami multimodalnymi w sektorze zdrowia. Dokument ten zawiera ponad czterdzieści precyzyjnych rekomendacji skierowanych do rządów, korporacji technologicznych oraz bezpośrednich dostawców usług medycznych, kładąc nacisk na odpowiedzialne wdrażanie algorytmów. WHO akcentuje, że modele te, zdolne do przetwarzania wielu typów danych i generowania różnorodnych wyników, otwierają bezprecedensowe możliwości w badaniach klinicznych i zdrowiu publicznym. Niemniej jednak ich potęga wymaga adekwatnego zarządzania, rygorystycznej oceny ryzyka oraz bezkompromisowej ochrony prywatności pacjentów, aby technologia służyła człowiekowi, a nie tylko statystyce.

Zjawisko określane jako „black box" budzi w medycynie szczególny niepokój, gdyż lekarz nie może abdykować z własnego osądu na rzecz prawdopodobieństwa, którego mechanizmów zupełnie nie rozumie. Wyjaśnialność, czyli zdolność systemu do przedstawienia logicznych przesłanek, które doprowadziły do podjęcia konkretnej decyzji, staje się w tym przypadku fundamentem bezpieczeństwa klinicznego. Jeżeli algorytm rozpoznaje zmianę nowotworową, specjalista musi znać czułość i swoistość systemu, a także wiedzieć, na jakich populacjach model był trenowany i gdzie zawodzi. Skuteczność bez sprawiedliwości jest bowiem tylko przyspieszoną niesprawiedliwością, zwłaszcza gdy system przewiduje przebieg choroby zakaźnej, a klinicysta musi traktować ten wynik jedynie jako wsparcie. Diagnostyka pozbawiona jasno zdefiniowanej odpowiedzialności byłaby w istocie medycznym odpowiednikiem sądu, w którym wyroki zapadają bez udziału sędziego i bez możliwości apelacji.

Właściwym kierunkiem rozwoju wydaje się formuła human-in-the-loop, czyli model współpracy, w którym człowiek pozostaje ostatecznym ogniwem decyzyjnym, co z powodzeniem testowano w Indiach podczas pandemii. Maszyna powinna wzmacniać percepcję klinicysty, lecz nigdy nie może zastępować jego etycznej i zawodowej odpowiedzialności za życie drugiego człowieka. Zdradliwość algorytmów medycznych polega na tym, że ich błąd bywa nierównomiernie rozłożony, co może prowadzić do ukrytych form dyskryminacji klinicznej. System może wykazywać imponującą skuteczność średnią, działając jednocześnie znacznie gorzej wobec kobiet, osób starszych czy mniejszości etnicznych spoza dominującego zbioru treningowego. To stary problem statystyki przebrany w nowoczesny, cyfrowy interfejs, gdzie ogólna suma wygląda zdrowo, podczas gdy konkretne podgrupy pacjentów metaforycznie krwawią.

Z tego powodu audyt medycznej sztucznej inteligencji musi być procesem stratyfikowanym, wrażliwym populacyjnie i połączonym z nieustannym monitoringiem już po wdrożeniu systemu do

pracy. Model, który okazał się bezpieczny i skuteczny w nowoczesnym szpitalu w Bostonie, wcale nie musi działać identycznie w powiatowej placówce w Polsce czy mobilnym punkcie w Afryce. Kwestia danych zdrowotnych odsłania przy tym bolesne napięcie między koniecznością innowacji a fundamentalnym prawem pacjenta do zachowania prywatności. Medycyna potrzebuje ogromnych zbiorów danych, by móc personalizować terapie, lecz pacjent potrzebuje ochrony informacji o swojej egzystencjalnej i cielesnej kruchości. Nie wolno nam akceptować narracji żądającej danych w zamian za obietnicę lepszego jutra, jeśli nie towarzyszą jej jasne odpowiedzi dotyczące anonimizacji i bezpieczeństwa. Przymierze z pacjentem bez pełnej przejrzystości jest bowiem tylko paternalizmem wyposażonym w nieco lepszy procesor.

W obszarze cyberbezpieczeństwa sztuczna inteligencja wprowadza zupełnie inną logikę, funkcjonując jednocześnie jako najtwardsza tarcza oraz wyjątkowo ostre i niebezpieczne ostrze. Po stronie obronnej systemy te potrafią błyskawicznie wykrywać anomalie, korelować sygnały z wielu źródeł oraz automatyzować odpowiedzi na incydenty, wspierając tym samym przeciążone zespoły specjalistów. W świecie, w którym liczba cyberataków dawno przekroczyła ludzką zdolność do manualnego reagowania, inteligentna automatyzacja staje się dla organizacji palącą koniecznością. Jednak po stronie agresora AI drastycznie obniża koszty phishingu, poprawia jakość socjotechniki i pomaga w masowym generowaniu złośliwego kodu. Deepfake, czyli syntetyczny materiał audiowizualny generowany przez AI, który w sposób realistyczny imituje rzeczywistość, pozwala na tworzenie personalizowanych oszustw o niespotykanej dotąd skali.

Raport ENISA dotyczący krajobrazu zagrożeń na rok 2025 wskazuje, że sztuczna inteligencja stała się jednym z definiujących elementów współczesnego pejzażu cyberzagrożeń. Dokument ten podkreśla niezwykle szybkie wykorzystywanie podatności systemowych po ich ujawnieniu, co sprawia, że przewaga czasowa obrońcy kurczy się w sposób wręcz dramatyczny. W tej sferze tradycyjne zarządzanie ryzykiem przestaje wystarczać, zwłaszcza gdy do gry wchodzą systemy agentowe zdolne do samodzielnego podejmowania działań technicznych. Taki agent może izolować hosty lub zmieniać reguły zapory, lecz źle zaprojektowany lub zaatakowany metodą prompt injection może stać się gorliwym sabotażystą. Istnieje realne ryzyko, że stworzymy strażnika, który w imię bezpieczeństwa zamknie wszystkich mieszkańców w piwnicy i ogłosi wielki sukces w redukcji włamań. Najlepszą metaforą nie jest tu zatem inteligentny asystent, lecz uzbrojony dyżurny nocny, który w panice może odciąć prąd w całym szpitalu.

Dlatego ramy operacyjne takie jak S.A.F.E. oraz rozwijane ujęcia TRUST, akcentujące transparentność i odporność, mają dziś znaczenie czysto praktyczne, a nie tylko retoryczne. Cyberobrona oparta na sztucznej inteligencji musi być projektowana w sposób umożliwiający odtworzenie każdej decyzji oraz natychmiastowe zatrzymanie autonomii systemu przez człowieka.

Konieczne jest wymuszenie eskalacji incydentów do operatora, ograniczenie zakresu uprawnień agentów oraz ciągłe monitorowanie ich zachowania w środowisku produkcyjnym. Przemysł 5.0 dodaje do tej technicznej dyskusji niezwykle istotny wymiar antropologiczny, przesuwając akcent z czystej automatyzacji ku głębokiej współpracy człowieka z maszyną. Podczas gdy poprzednia rewolucja koncentrowała się na optymalizacji procesów i Internecie Rzeczy, obecna faza ma stawiać w centrum odporność, zrównoważenie oraz dobrostan pracownika.

To szlachetne podejście wymaga jednak szczególnej czujności, gdyż każde hasło o „ludzkocentryczności" może stać się jedynie pustą dekoracją, jeśli człowiek pozostanie biologicznym dodatkiem do systemu. Kluczowe pytanie brzmi, czy sztuczna inteligencja będzie realnie wzmacniać kompetencje pracowników, czy raczej zamieni ich w monitorowane sensory, których ciało optymalizuje się jak parametry linii. W nowoczesnych zakładach AI może wspierać bezpieczeństwo i przewidywać awarie, ale równie dobrze może służyć do intensyfikacji kontroli i redukcji zawodowej autonomii. Etyka technologii nierozerwalnie łączy się tutaj z socjologią pracy, wymagając ochrony godności jednostki w środowisku, które dąży do totalnej przejrzystości. Pracownik musi zachować prawo do zrozumienia kryteriów swojej oceny oraz możliwość zakwestionowania wyniku wygenerowanego przez algorytm zarządzający. Fabryka przyszłości nie może stać się cyfrowym panoptikonem, w którym produktywność osiąga się kosztem uprzedmiotowienia człowieka przez robotyczne ramię. Etyka nie może być przecież kwiatkiem przypiętym do serwera, lecz musi stanowić integralną część kodu, na którym budujemy naszą przyszłość.

Etyka w praktyce: od asfaltu po ławki szkolne

Transport autonomiczny to miejsce, gdzie abstrakcyjna etyka brutalnie zderza się z twardym asfaltem rzeczywistości. Pojazd, pozbawiony ludzkiego instynktu, musi podejmować krytyczne decyzje w warunkach permanentnej niepewności oraz drastycznie ograniczonego czasu reakcji. Media chętnie spłycają ten problem do dylematu wagonika, czyli klasycznego problemu etycznego dotyczącego wyboru między dwoma tragicznymi w skutkach scenariuszami, jakby samochody miały regularnie rozstrzygać o życiu pasażerów i pieszych. W rzeczywistości jednak wyzwania mają charakter znacznie bardziej prozaiczny, inżynieryjny oraz proceduralny. Kluczowe staje się to, jak system rozpoznaje obiekty, przewiduje trajektorie i reaguje na awarie czujników w trybie minimalnego ryzyka. Etyka nie może być kwiatkiem przypiętym do serwera; musi zostać wpisana w samą definicję bezpieczeństwa oraz dopuszczalnego ryzyka na drodze.

Obecnie UNECE prowadzi intensywne prace nad regulacjami dla pojazdów zautomatyzowanych, dążąc do wypracowania globalnych standardów zarządzania bezpieczeństwem w ruchu lądowym. Dokumenty grup roboczych szczegółowo opisują Automated Driving Systems, czyli zestaw środków projektowych mających zapewnić, że system działa bez nieuzasadnionego ryzyka dla otoczenia. Zapowiadane na 2026 rok regulacje oznaczają, że transport autonomiczny ostatecznie opuszcza fazę radosnego eksperymentu technologicznego na rzecz międzynarodowego nadzoru. W tym procesie kluczowe okazuje się pojęcie translational knowledge, czyli umiejętność przełożenia specjalistycznej wiedzy między różnymi, często odległymi od siebie dziedzinami. Etyk, inżynier i prawnik muszą wreszcie wypracować wspólny język, aby wartości moralne przestały być jedynie teoretycznym konstruktem. Wartości te należy bowiem przetłumaczyć na konkretne scenariusze, progi czułości i mechanizmy awaryjne, inaczej etyka pozostanie jedynie pięknym esejem napisanym już po wypadku.

Administracja publiczna oraz wymiar sprawiedliwości to obszary, w których sztuczna inteligencja dotyka samej istoty relacji obywatela z aparatem państwowym. Państwo, dysponując monopolem na legalny przymus i rozdzielając kluczowe świadczenia, musi spełniać standardy etyczne znacznie wyższe niż jakakolwiek prywatna korporacja. Podczas gdy firma może po prostu wycofać wadliwy produkt z rynku, państwo nie naprawi tak łatwo wyrządzonej krzywdy systemowej. Niesłusznie utracony czas, zniszczona reputacja czy odebrana wolność to koszty, których nie da się zniwelować prostą aktualizacją oprogramowania w chmurze. Algorytmiczna administracja kusi oczywiście obietnicą niespotykanej efektywności, szybszego rozpatrywania wniosków oraz precyzyjnego wykrywania nadużyć finansowych. Niemniej jednak skuteczność bez sprawiedliwości jest tylko przyspieszoną niesprawiedliwością.

Cyfrowe państwo pozbawione jasnych zasad może niepostrzeżenie przeobrazić się w system permanentnego podejrzenia wobec własnych obywateli. Jeżeli algorytm typuje jednostkę do kontroli skarbowej, musi ona wiedzieć, na jakiej konkretnie podstawie podjęto taką decyzję i jak może ją zakwestionować. Tutaj kluczowa staje się wyjaśnialność (XAI), czyli zdolność systemu do przedstawienia logicznych przesłanek, które doprowadziły do podjęcia konkretnej decyzji, co pozwala na realną kontrolę nad algorytmami. W dojrzałej demokracji każda decyzja publiczna musi mieć twarz konkretnej odpowiedzialności, nawet jeśli proces obliczeniowy wykonała maszyna. Wolność obywatelska rzadko bywa odbierana w akompaniamencie wagnerowskich trąb; znacznie częściej wyparowuje ona po cichu, podczas bezrefleksyjnego przeglądania kolejnych ekranów. Bez transparentności ryzykujemy budowę biurokracji, w której uzasadnienie „model tak ocenił” staje się ostatecznym i niepodważalnym wyrokiem.

Kolejnym wyzwaniem jest ryzyko pogłębienia wykluczenia cyfrowego, które może stać się nową formą instytucjonalnego analfabetyzmu przerzuconego na najsłabszych. AI teoretycznie poprawia dostępność usług dla osób z niepełnosprawnościami czy mieszkańców peryferii, oferując im wsparcie w wielu językach jednocześnie. Jednakże ta sama technologia może stać się barierą nie do przebycia dla tych, którym brakuje kompetencji cyfrowych lub stabilnego łącza internetowego. Państwo nie ma prawa karać obywatela za to, że nie potrafił on sformułować odpowiednio skutecznego zapytania do urzędowego chatbota. Przerzucanie ciężaru obsługi skomplikowanych procedur na obywatela jest nie tylko nieuczciwe, ale wręcz systemowo szkodliwe. Technologia powinna budować mosty, a nie wznosić mury, których sforsowanie wymaga posiadania najnowszego smartfona i biegłości w kodowaniu.

Sektor edukacji ujawnia podobne napięcia, choć przybierają one formę znacznie bardziej subtelną i rozłożoną na całe pokolenia. Sztuczna inteligencja może być genialnym korepetytorem lub laboratorium krytycznego myślenia, wspierającym uczniów o specjalnych potrzebach edukacyjnych w ich rozwoju. Istnieje jednak realne ryzyko utrwalenia tak zwanego bankowego modelu edukacji, w którym uczeń jedynie biernie gromadzi gotowe odpowiedzi dostarczone przez algorytm. Jeżeli młody człowiek używa technologii do zastąpienia wysiłku intelektualnego, zamiast go pogłębiać, to proces emancypacji zostaje brutalnie przerwany. Maszyna zdejmuje z nas ciężar myślenia, ale przy okazji niepostrzeżenie odbiera nam mięsień sprawczości i zdolność do samodzielnego problematyzowania świata. Technologia, niczym mityczny ogień, daje nam potęgę, ale wymaga też pokory wobec jej niszczycielskiej siły.

W tym kontekście prompt engineering, czyli technika precyzyjnego formułowania zapytań kierowanych do modelu, może być albo tanią sztuczką, albo głębokim ćwiczeniem epistemologicznym. W wersji dojrzałej zmusza on człowieka do precyzowania założeń, rozpoznawania ukrytych uprzedzeń i krytycznej analizy otrzymanych od modelu wariantów. Użytkownik świadomy nie pyta maszyny po to, aby szybko zakończyć proces myślowy i otrzymać gotowy produkt do wklejenia. Pyta raczej po to, by skomplikować własne rozumowanie i skonfrontować je z perspektywą sceptyka oraz adwokata diabła naraz. To zasadnicza różnica między uczniem kopiującym wynik a badaczem prowadzącym fascynujący dialog z nieskończoną, choć czasem omylną biblioteką. W tym punkcie idee Michela Foucaulta spotykają się z myślą Paulo Freirego, definiując na nowo granice współczesnego poznania.

Systemy AI w szkołach mogą jednak łatwo stać się technologiami dyscyplinowania, które drobiazgowo mierzą każdą sekundę aktywności i reakcji ucznia. Z jednej strony pozwala to na personalizację nauczania, z drugiej zaś sprowadza żywego człowieka do zestawu suchych danych statystycznych i predykcji. Jeżeli szkoła przyszłości stanie się jedynie adaptacyjną platformą

optymalizującą testowalne kompetencje, wytworzy ona posłuszne ciała, a nie wolnych i krytycznych obywateli. Aluzja do panoptikonu, czyli modelu więzienia umożliwiającego stałą obserwację, nie jest tutaj jedynie literacką ozdobą, lecz realnym ostrzeżeniem przed nadzorem. Musimy zadbać, aby spojrzenie algorytmu nie stało się dla ucznia ważniejsze niż autentyczny, pełen błędów i poszukiwań dialog z nauczycielem. Etyka w edukacji nie może być tylko przypisem w regulaminie, lecz fundamentem, na którym budujemy naszą wspólną przyszłość.

Demokracja w erze AI: między innowacją a odpowiedzialnością

Polityczny wymiar sztucznej inteligencji nie jest jedynie teoretycznym konstruktem, lecz manifestuje się najpełniej w samej tkance współczesnej demokracji, stając się jej nową, cyfrową infrastrukturą uczestnictwa. AI może przecież wspierać kampanie wyborcze poprzez precyzyjne tłumaczenie treści na lokalne dialekty, syntetyzowanie programów partyjnych dla zabieganego obywatela czy analizowanie nastrojów społecznych z precyzją, o której dawni socjolodzy mogli tylko marzyć. Algorytmy potrafią zwiększać dostępność debat publicznych i pomagać wyborcom w merytorycznym porównywaniu propozycji, co teoretycznie powinno wzmacniać fundamenty republiki. Niemniej ta sama technologia, niczym mityczny Janus, posiada drugie, znacznie mroczniejsze oblicze, które zagraża zaufaniu społecznemu poprzez generowanie deepfake'ów, czyli syntetycznych materiałów audiowizualnych, które w sposób przerażająco realistyczny imitują rzeczywistość. Obywatel nie traci wolności w dramatycznej scenie z muzyką Wagnera; traci ją, scrollując strumień spreparowanych obrazów, fałszywych poparć i mikrotargetowanej propagandy, która żeruje na jego najgłębszych lękach.

Najbardziej perfidnym owocem generatywnej sztucznej inteligencji okazuje się tak zwana dywidenda kłamcy, zjawisko, w którym sama obecność technologii fałszowania staje się tarczą dla osób faktycznie winnych. Polityk przyłapany na autentycznym, kompromitującym nagraniu nie musi już udowadniać swojej niewinności; wystarczy, że rzuci cień podejrzenia, sugerując, iż materiał jest produktem algorytmu. To przesuwa ciężar dowodu z prawdy na wszechobecną niepewność, co jest strategią niezwykle skuteczną w spolaryzowanym społeczeństwie, gdzie emocje dawno wygrały z faktami. Autokrata nie potrzebuje już przekonywać mas do swojej wersji wydarzeń, gdyż wystarczy mu, by obywatele zwątpili w istnienie jakiejkolwiek obiektywnej wersji rzeczywistości. W takim układzie sił prawda przestaje być punktem odniesienia, a staje się jedynie jedną z wielu opcji w menu dezinformacji, co nieuchronnie prowadzi do erozji fundamentów

demokratycznego dialogu. To nie jest postęp, to jest ucieczka od odpowiedzialności w cyfrową chmurę.

Kiedy cynizm zastępuje krytycyzm, staje się on swoistym politycznym środkiem znieczulającym, który sprawia, że obywatel przestaje wierzyć, ale jednocześnie przestaje działać. W nowoczesnej demokracji problemem nie jest już wyłącznie dezinformacja rozumiana jako pojedyncze kłamstwo, lecz cały dezinformacyjny porządek, czyli środowisko, w którym prawda traci swoją instytucjonalną przewagę nad fałszem. Dawniej propaganda musiała mozolnie produkować silne przekonania, by mobilizować masy do określonych celów, dziś natomiast często wystarczy produkować zmęczenie i poczucie przytłoczenia nadmiarem sprzecznych bodźców. Obywatel bombardowany syntetycznymi obrazami, ekspertami znikąd i emocjonalnymi komunikatami zaczyna w końcu machać ręką, uznając, że wszyscy kłamią. A w świecie, w którym wszyscy kłamią, ostateczne zwycięstwo odnosi ten, kto najsprawniej zarządza kłamstwem i potrafi najskuteczniej administrować społecznym gniewem.

Dlatego też znakowanie treści syntetycznych oraz cyfrowe poświadczenia pochodzenia są dziś palącą koniecznością, choć musimy mieć świadomość, że pozostają one rozwiązaniami dalece niewystarczającymi. Etykiety typu „AI Info" mogą wprawdzie pomóc użytkownikowi rozpoznać materiał generowany, ale nie rozwiążą one głębokiego kryzysu zaufania, który toczy nasze społeczeństwa. Po pierwsze, profesjonalni fałszerze będą z oczywistych względów unikać znakowania, a po drugie, samo oznaczenie może zostać wykorzystane strategicznie do podważania treści całkowicie prawdziwych. Co więcej, użytkownik przeciążony informacyjnie może zacząć ignorować te etykiety z taką samą obojętnością, z jaką akceptuje kolejne regulaminy usług cyfrowych. Potrzebujemy zatem nie tylko technicznych znaczników, lecz przede wszystkim rzetelnej edukacji medialnej, niezależnych instytucji weryfikacyjnych oraz kultury politycznej, która przestanie nagradzać cyniczne kłamstwo mianem „skutecznego przekazu".

Czas jednak podkreślić wątpliwość, której uczciwy esej nie powinien ukrywać przed czytelnikiem: nie mamy jeszcze pełnej, stabilnej wiedzy empirycznej o skali wpływu generatywnej AI na wyniki wyborów. Choć mnożą się przykłady incydentów, kampanii botów i deepfake'ów, to precyzyjne przełożenie tych zjawisk na konkretne decyzje wyborcze pozostaje niezwykle trudne do zmierzenia. Badacze słusznie ostrzegają przed paniką moralną, która każdą nieoczekiwaną zmianę polityczną czy sukces populistów próbuje tłumaczyć wszechmocnym „algorytmem". Jednocześnie brak twardego dowodu na bezpośredni wpływ nie jest bynajmniej dowodem na brak ryzyka, gdyż demokracja jest systemem wrażliwym nie tylko na ostateczny wynik, ale przede wszystkim na zaufanie do samej procedury. Nawet jeśli AI nie zmieniła jeszcze wyniku żadnych wyborów, może

skutecznie zmienić sposób, w jaki obywatele interpretują prawomocność całego procesu demokratycznego.

W sektorach krytycznych musimy zatem nauczyć się rozróżniać między ryzykiem udokumentowanym, prawdopodobnym a czysto spekulatywnym, unikając popadania w skrajności. Nie każda sztuczna inteligencja dyskryminuje, nie każdy chatbot halucynuje w sposób szkodliwy i nie każda regulacja automatycznie poprawi nasze bezpieczeństwo. Sceptycyzm wobec alarmizmu nie może jednak przerodzić się w infantylny optymizm, który ignoruje realne zagrożenia płynące z niekontrolowanego rozwoju technologii. Właściwa postawa powinna być inżynieryjno-republikańska: należy zakładać możliwość błędu, projektować mechanizmy kontroli, monitorować skutki i utrzymywać człowieka jako ostateczne ogniwo odpowiedzialne za system. Tylko poprzez ciągłą aktualizację ram nadzoru możemy nadążyć za tempem zmian, które sami sprowokowaliśmy.

W tym kontekście pojawia się AI Act, czyli europejskie rozporządzenie o sztucznej inteligencji wprowadzające ramy prawne dla bezpiecznego rozwoju technologii, które staje się manifestem cywilizacyjnym starego kontynentu. Krytycy z kręgów techno-libertariańskich i akceleracjonistycznych grzmią, że nadmierna ostrożność zabije innowację, spowolni rozwój medycyny i odda przewagę państwom mniej skrupulatnym w kwestiach etycznych. Warto potraktować tę krytykę poważnie, ponieważ etyka AI nie może stać się jedynie biurokratyczną liturgią produkowania polityk, których nikt nigdy nie przeczyta, a które jedynie duszą małych innowatorów. Z tego ostrzeżenia nie wynika jednak, że należy porzucić próby uregulowania tego obszaru, lecz raczej, że regulacja musi być inteligentna, proporcjonalna i technicznie kompetentna. Argument głoszący, że „regulacja zabija innowację", jest często zbyt wygodnym wytrychem; równie dobrze można by twierdzić, że normy bezpieczeństwa lotniczego zabiły marzenie o lataniu.

W rzeczywistości to właśnie mądre reguły gry tworzą zaufanie, a zaufanie jest jedynym paliwem, które pozwala na szeroką i trwałą adopcję nowych technologii w społeczeństwie. Nikt rozsądny nie wsiądzie do autonomicznego pojazdu tylko dlatego, że startup ma piękne logo, i nikt nie odda swoich danych zdrowotnych systemowi, który opiera swój autorytet wyłącznie na obietnicy bycia „przełomowym". Innowacja bez zaufania jest jedynie pokazem pirotechnicznym w magazynie benzyny, który może efektownie wyglądać, ale grozi katastrofą przy najmniejszym błędzie. Z drugiej strony barykady stoją krytycy radykalni, twierdzący, że etyczna AI to oksymoron, gdyż systemy te z natury służą ekstrakcji danych i koncentracji kapitału. Ich ostrzeżenie jest cenne: nie wolno nam redukować problemu AI do prostej listy wartości, jeśli struktura władzy i dostępu do infrastruktury obliczeniowej pozostaje nietknięta w rękach nielicznych gigantów.

Esej ten opowiada się za stanowiskiem źródłowym: odpowiedzią na ryzyko nie jest rezygnacja z postępu, lecz budowa systemów transparentnych, audytowalnych i społecznie osadzonych. Krytyka

strukturalna jest konieczna, ale nie może ona paraliżować działań tam, gdzie technologia może przynieść realną ulgę w cierpieniu czy poprawę bezpieczeństwa. Jeżeli lekarz może dzięki algorytmom szybciej wykryć chorobę, nie wolno mu powiedzieć pacjentowi, by poczekał, aż najpierw rozwiążemy wszystkie problemy kapitalizmu platformowego. Dojrzała odpowiedź musi być zatem podwójna: musimy wdrażać technologię tam, gdzie służy ona dobru wspólnemu, i jednocześnie budować instytucje zdolne do ograniczania jej nadużyć. Prometeusz nie musi oddawać ognia, który podarował ludzkości, ale musi wreszcie przestać udawać, że ogień ten nie parzy i nie wymaga ostrożnego obchodzenia się z płomieniem.

Zastosowania AI w sektorach krytycznych pokazują nam jedną wspólną zasadę: nie ma odpowiedzialnej automatyzacji bez odpowiedzialnej instytucji, która za nią stoi. Żaden model, choćby najbardziej zaawansowany, nie zastąpi kultury organizacyjnej, a żaden audyt nie zastąpi odwagi menedżerskiej w sytuacjach granicznych. Wyjaśnialność, czyli zdolność systemu do przedstawienia logicznych przesłanek, które doprowadziły do podjęcia konkretnej decyzji, nie zastąpi nigdy prawa obywatela do sprzeciwu i ludzkiego osądu. Koncepcja human-in-the-loop, zakładająca udział człowieka w pętli decyzyjnej, ma sens tylko wtedy, gdy ten człowiek posiada realną władzę zatrzymania systemu, a nie jest jedynie figurantem podpisującym algorytmiczne wyroki. Ostatecznie żadna infrastruktura danych nie zastąpi fundamentalnego pytania Thoreau: nad czym tak pracowicie się trudzisz i komu ta praca w rzeczywistości służy? Etyka nie może być przecież kwiatkiem przypiętym do serwera, lecz musi stanowić sam rdzeń kodu, na którym budujemy naszą wspólną przyszłość.

> Technologia jest jedynie lustrem, w którym odbijają się nasze własne słabości i dążenia do sprawczości. Pytanie o przyszłość AI nie dotyczy tego, jak szybko maszyny zaczną myśleć, lecz czy my sami zachowamy zdolność do rozumnego wyboru w świecie pełnym algorytmicznych podpowiedzi. Czy staniemy się architektami nowej sprawiedliwości, czy jedynie nieświadomymi zakładnikami systemów, które stworzyliśmy w imię własnej wygody?

Inspiracje

[1] **"Ethical AI and Data Science"** Tereza Raquel Merlo, Jay Liebowitz

2026 | Routledge | ISBN: 9781040651780

Tereza Raquel Merlo jest badaczką podoktorską na Federalnym Uniwersytecie Rio de Janeiro (UFRJ), afiliowaną przy Brazylijskim Instytucie Informacji w Nauce i Technologii (IBICT), oraz adiunktem na University of North Texas. Jej główne obszary działalności obejmują zarządzanie wiedzą, sztuczną inteligencję i analizę danych. Jest autorką książki „Understanding, Implementing, and Evaluating Knowledge Management in Business Settings" (2022) oraz redaktorką „Ethical AI and Data Science" (2026).

Tereza Raquel Merlo *is a postdoctoral researcher at the Federal University of Rio de Janeiro (UFRJ), affiliated with the Brazilian Institute of Information in Science and Technology (IBICT), and an Adjunct Professor at the University of North Texas. Her main areas of activity include knowledge management, artificial intelligence, and data analytics. She is the author of "Understanding, Implementing, and Evaluating Knowledge Management in Business Settings" (2022) and editor of "Ethical AI and Data Science" (2026).*

Jay Liebowitz to wybitny autorytet w dziedzinie sztucznej inteligencji (AI) i autor licznych publikacji. Jest profesorem innowacji biznesowych i dyrektorem Centrum AI-EDGE w Crummer Graduate School of Business na Rollins College. Wcześniej pełnił funkcje m.in. w NASA oraz na Johns Hopkins University. Jest również założycielem i redaktorem naczelnym czasopisma "Expert Systems With Applications".

Dr. ***Jay Liebowitz*** *is a prominent scholar and practitioner in the fields of knowledge management, data analytics, and artificial intelligence. Currently serving as a Special Advisor to the President at Thomas Jefferson University, he has held numerous prestigious academic and leadership positions, including Executive-in-Residence for Public Service at Columbia University's Data Science Institute and Professor of Business Innovation and Industry Transformation at Rollins College. Dr. Liebowitz is the founding editor-in-chief of the top-tier journal 'Expert Systems with Applications' and has authored or edited over 50 books. His work focuses on the intersection of technology, organizational intelligence, and ethics, particularly in how AI and data science can be applied responsibly in government, business, and education. He is recognized globally for his contributions to intelligent systems, IT management, and the development of intuition-based decision-making frameworks in the era of digital transformation.*

Thomas Jefferson University

Literatura uzupełniająca

Książki

Literatura pokrewna (Google Scholar):

[1] "Walden; or, Life in the Woods"

Henry David Thoreau

1854 | Ticknor and Fields | ISBN: 978-0140390445

[2] "A Week on the Concord and Merrimack Rivers"

Henry David Thoreau

1849 | James Munroe and Company | ISBN: 978-0691065090

[3] "Etyka nikomachejska"

Arystoteles

-350 | Wydawnictwa klasyczne (różne wydania) | ISBN: 978-83-01-14935-2

[4] "Polityka"

Arystoteles

-330 | Wydawnictwa klasyczne (różne wydania) | ISBN: 978-83-01-14936-9

[5] "The Republic"

Plato

-375 | N/A (Ancient Text)

[6] "Symposium"

Plato

-385 | N/A (Ancient Text)

[7] "Opera and Drama"

Richard Wagner

1851 | J.J. Weber | ISBN: 978-0803297654

[8] "The Artwork of the Future"

Richard Wagner

1849 | Otto Wigand | ISBN: 978-0803297524

[9] "Dune"

Frank Herbert

1965 | Chilton Books | ISBN: 978-0-441-17271-9

[10] "Destination: Void"

Frank Herbert

1966 | Berkley Books | ISBN: 978-0-425-08638-4

[11] "Public Opinion"

Walter Lippmann

1922 | Harcourt, Brace and Company | ISBN: 978-0-02-919130-6

[12] "The Phantom Public"

Walter Lippmann

1925 | Harcourt, Brace and Company | ISBN: 978-1-56000-677-0

[13] "Pedagogy of the Oppressed"

Paulo Freire

1970 | Continuum | ISBN: 978-0-8264-1276-8

[14] "Cultural Action for Freedom"

Paulo Freire

1970 | Harvard Educational Review | ISBN: 978-0-916690-37-3

[15] "Discipline and Punish: The Birth of the Prison"

Michel Foucault

1975 | Gallimard (French), Vintage Books (English) | ISBN: 978-0-679-75255-4

[16] "The History of Sexuality, Vol. 1: An Introduction"

Michel Foucault

1976 | Gallimard (French), Pantheon (English) | ISBN: 978-0-679-72469-8

[17] "Ethical AI and Data Science: Building Trustworthy and Transparent Systems"

Tereza Raquel Merlo, Jay Liebowitz

2026 | Auerbach Publications / Routledge | ISBN: 978-1-0411-0929-7

[18] "Pivoting Government Through Digital Transformation"

Jay Liebowitz

2024 | Taylor & Francis | ISBN: 978-1-0326-5573-4

[19] **"The Burnout Society"**
Byung-Chul Han
2010 | Stanford University Press

[20] **"The Ethics of Artificial Intelligence"**
S. Matthew Liao
2020 | Oxford University Press

[21] **"Creating Capabilities: The Human Development Approach"**
Martha Nussbaum
2011 | Harvard University Press

[22] **"The Shallows: What the Internet Is Doing to Our Brains"**
Nicholas Carr
2010 | W. W. Norton & Company

[23] **"Human Compatible: Artificial Intelligence and the Problem of Control"**
Stuart Russell
2019 | Viking

[24] **"The Public and Its Problems"**
John Dewey
2012 | Ohio University Press

[25] **"The Social Philosophers"**
Robert Nisbet
2025 | American Philosophical Society Press

[26] **"The Idea of Justice"**
Amartya Sen
2009 | Harvard University Press

[27] **"The Righteous Mind: Why Good People Are Divided by Politics and Religion"**
Jonathan Haidt
2012 | Pantheon

Artykuły naukowe

Autorzy przywoływani:

[1] "The Relevance of Philosophy in Shaping Contemporary Social Ethics"

Jakoep Ezra Harianto

2025 | International Journal for Science Review

Google Scholar

[2] "Civil Disobedience (Resistance to Civil Government)"

Henry David Thoreau

1849 | Aesthetic Papers

Google Scholar

[3] "Virtue Ethics and Artificial Intelligence"

Shannon Vallor

2016 | Philosophy & Technology

Google Scholar

[4] "MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias"

F. Barsotti, R.G. Koçer

2024 | AI & SOCIETY

Google Scholar

[5] "Judaism in Music (Das Judenthum in der Musik)"

Richard Wagner

1850 | Neue Zeitschrift für Musik

Google Scholar

[6] "Beethoven"

Richard Wagner

1870 | E.W. Fritzsch

Google Scholar

[7] "The Ethics of Digital Manipulation"

Cass R. Sunstein

2016 | Journal of Political Philosophy

Google Scholar

[8] "Humans, Machines, and an Ethics for Technology in Dune"

Zachary Pirtle

2023 | Dune and Philosophy: Minds, Monads, and Muad'dib (Wiley)

Google Scholar

[9] "The Scholar in a Troubled World"

Walter Lippmann

1932 | The Atlantic Monthly

Google Scholar

[10] "Walter Lippmann and Public Opinion"

Tom Arnold-Forster

2023 | History of Political Thought

Google Scholar

[11] "Justice for Girls: On the Provision of Abortion as Adequate Care"

Anonymous

2026 | Ethics

Google Scholar

[12] "The Subject and Power"

Michel Foucault

1982 | Critical Inquiry | DOI: 10.1086/448181

Google Scholar

[13] "What is Critique?"

Michel Foucault

1978 | Bulletin de la Société française de philosophie

Google Scholar

Tekst opracowany z udziałem sztucznej inteligencji.
Redakcja i kontrola merytoryczna: Fundacja Dobre Państwo.
APA ONE Publishing System

Article ID: b3c08b90-2e78-4b2a-b444-ef69a52da1e0