Automatyzacja sądu i kompetencji: etyka, prawo i przyszłość pracy w dobie AI według Brada Smitha

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje przejście od automatyzacji prostych czynności do „automatyzacji sądu", gdzie AI współuczestniczy w kluczowych klasyfikacjach społecznych i prawnych. Autor podkreśla, że systemy ML dostarczają jedynie sygnałów probabilistycznych, a nie gotowe decyzje. Kluczowym problemem jest przejście od predykcji do normy – ustalenie progu decyzyjnego jest wyborem wartościującym, a nie technicznym. Tekst ostrzega przed złudzeniem obiektywizmu AI, wskazując na ryzyko błędów wynikających z niewłaściwych zmiennych proxy (np. koszt leczenia jako miernik zdrowia). Wnioski prowadzą do tezy, że sprawiedliwość algorytmiczna wymaga świadomego nadzoru ludzkiego i korekty normatywnej, gdyż modele mogą wiernie odzwierciedlać i utrwalać istniejące w świecie nierówności.

8. The article analyzes the transition from automating simple tasks to the "automation of judgment," where AI participates in key social and legal classifications. The author emphasizes that ML systems provide only probabilistic signals, not ready-made decisions. The core problem is the transition from prediction to norm—setting a decision threshold is a value-based choice, not a technical one. The text warns against the illusion of AI objectivity, pointing to the risk of errors resulting from inappropriate proxy variables (e.g., treatment cost as a measure of health). The conclusions lead to the thesis that algorithmic fairness requires conscious human oversight and normative correction, as models can faithfully reflect and perpetuate existing inequalities in the world.

Artykuł analizuje fundamentalną zmianę paradygmatu technologicznego: przejście od automatyzacji prostych czynności do automatyzacji ludzkiego sądu. Autor stawia tezę, że

delegowanie decyzji o charakterze normatywnym (np. w medycynie, prawie czy finansach) do systemów probabilistycznych nie jest problemem technicznym, lecz głęboko politycznym i etycznym, gdyż każdy algorytm koduje określone wartości i ryzyka. W obliczu zjawisk takich jak 'automation bias' oraz erozji klasy średniej, tekst postuluje konieczność przekształcenia miękkich postulatów etycznych w twarde ramy prawne (np. AI Act) oraz instytucjonalną infrastrukturę odpowiedzialności. Kluczowym elementem tej strategii jest przebudowa edukacji: szkoła musi przestać być dostarczycielem informacji, a stać się gwarantem bezpieczeństwa poznawczego. Tylko poprzez rozwój metakognicji i zachowanie wiedzy faktograficznej jako rusztowania dla krytycznego myślenia, człowiek może uniknąć intelektualnej zależności od maszyn i zachować suwerenność w świecie, w którym granica między kompetencją a jej stylistyczną symulacją ulega zatarciu.

SŁOWA KLUCZOWE

automatyzacja sądu | odpowiedzialna AI | algorytmiczna sprawiedliwość | automation bias | human-in-the-loop | distribution shift | calibrated trust | suwerenność epistemiczna | błąd perswazyjny | meaningful human control | NIST AI Risk Management Framework | metakognicja | proxy zdrowia | fairness w ML | wyjaśnialność normatywna

Przejście od automatyzacji czynności do automatyzacji sądu

Od automatyzacji działania do automatyzacji sądu. Sztuczna inteligencja, błąd, sprawiedliwość i granice delegowania decyzji

Dotychczas algorytmy decydowały przede wszystkim o tym, co człowiek zobaczy. Wraz z rozwojem sztucznej inteligencji przesuwamy się jednak ku znacznie głębszej formie delegowania: systemy zaczynają współuczestniczyć w rozstrzyganiu, co jest podejrzane, chore, niebezpieczne, wiarygodne czy wartościowe, a także kto zasługuje na kredyt, wymaga interwencji lub dalszej uwagi. Nie jest to już jedynie automatyzacja czynności, lecz automatyzacja klasyfikacji, prognozy i selekcji – funkcji, które przez stulecia należały do domeny ludzkiego sądu.

Smith ujmuje tę zmianę poprzez sześć zasad odpowiedzialnej AI: sprawiedliwość, niezawodność i bezpieczeństwo, prywatność i cyberbezpieczeństwo, inkluzywność, przejrzystość oraz odpowiedzialność. Ich znaczenie staje się jednak widoczne dopiero wtedy, gdy przestaniemy traktować je jako kodeks korporacyjnych cnót, a zaczniemy analizować jako odpowiedź na głęboki problem epistemologiczny: co właściwie dzieje się, gdy społeczeństwo zaczyna działać na podstawie prawdopodobieństw produkowanych przez modele?

Klasyczna maszyna automatyczna wykonywała instrukcję. Termostat reagował na temperaturę, kalkulator realizował działanie matematyczne, a program księgowy stosował zapisane reguły. System uczenia maszynowego działa inaczej: na podstawie danych historycznych konstruuje funkcję pozwalającą oszacować nieznany wynik dla nowego przypadku. Z punktu widzenia statystyki model nie stwierdza zatem, że „ten człowiek dopuści się przestępstwa", „ta pacjentka jest chora" czy „ten kandydat okaże się złym pracownikiem". Mówi raczej, że w przestrzeni cech poznanych z danych przypadek ten należy do obszaru, któremu model przypisuje określone prawdopodobieństwo wyniku.

Dopiero instytucja przekształca probabilistyczny sygnał w decyzję. Jest to kluczowe rozróżnienie, ponieważ społeczne użycie AI wymaga wykonania dodatkowego, często niewidzialnego kroku: przejścia od predykcji do normy. Model może oszacować ryzyko na poziomie 0,63, ale sam wynik nie odpowiada na pytanie, czy należy odmówić kredytu, skierować pacjenta na dodatkowe badania, zatrzymać podróżnego, przyznać świadczenie albo zwiększyć nadzór. Ktoś musi ustalić próg, a każdy próg oznacza wybór pomiędzy kosztami różnych rodzajów błędu.

Obniżenie progu może zwiększyć wykrywalność, ale jednocześnie podniesie liczbę fałszywych alarmów. Podwyższenie go ograniczy pomyłki, lecz pozostawi więcej przypadków niewykrytych. Matematyka potrafi policzyć konsekwencje, ale nie może sama zdecydować, który błąd społeczeństwo uznaje za bardziej dopuszczalny. W medycynie wynik fałszywie ujemny może oznaczać przeoczenie choroby, podczas gdy fałszywie dodatni prowadzi do lęku i ryzyka niepotrzebnego leczenia. W obszarze bezpieczeństwa fałszywy negatyw może przepuścić zagrożenie, a fałszywy pozytyw skierować niewinną osobę do kontroli. W wymiarze sprawiedliwości zaś zawyżona ocena ryzyka może prowadzić do ograniczenia wolności, a zaniżona – zwiększyć niebezpieczeństwo dla innych.

Nie istnieje więc „neutralny próg algorytmiczny". Każdy próg jest zakodowaną relacją pomiędzy wartościami. Tu zaczyna się pierwsza głęboka korekta wobec języka „obiektywnej AI". Model może być matematycznie precyzyjny, ale jego zastosowanie nie jest wolne od normatywności. Już samo określenie zmiennej celu wymaga decyzji o tym, co system ma przewidywać. Jeśli chcemy zidentyfikować „dobrego pracownika", musimy wcześniej zdefiniować dobro pracy. Przewidując

„ryzyko pacjenta”, trzeba wskazać, czy chodzi o śmiertelność, hospitalizację, koszt leczenia czy konkretne powikłanie. Jeśli system ma wykrywać „podejrzane zachowanie”, ktoś musi zdecydować, co uznajemy za normę, a co za odchylenie.

W tym momencie dane przestają być surowcem, a stają się teorią świata zapisaną w zmiennych. Zmienna zależna, etykieta, proxy i kryterium sukcesu nie są jedynie elementami technicznymi, lecz założeniami epistemologicznymi. Doskonałym przykładem tego problemu jest badanie Ziada Obermeyera i współautorów dotyczące algorytmu stosowanego w amerykańskiej ochronie zdrowia. System wykorzystywał przyszłe koszty opieki jako przybliżenie potrzeb zdrowotnych.

Okazało się jednak, że przy tym samym wyniku ryzyka czarni pacjenci byli przeciętnie bardziej chorzy niż biali. Źródłem problemu nie było jawne wprowadzenie rasy jako kryterium, lecz wybór kosztu jako proxy zdrowia w społeczeństwie, w którym dostęp do opieki i wydatki medyczne są nierówno rozłożone. Po skorygowaniu celu odsetek czarnych pacjentów kwalifikowanych do dodatkowej pomocy wzrósł w analizie autorów z 17,7 do 46,5 procent.

Algorytmiczna sprawiedliwość to wybór normatywny, a nie problem techniczny

Przykład ten ma znaczenie wykraczające daleko poza medycynę. Pokazuje, że algorytm może wiernie nauczyć się niesprawiedliwego świata, nie popełniając przy tym błędu statystycznego w banalnym sensie. Jeśli historyczny system ekonomiczny, administracyjny czy społeczny traktował grupy nierówno, dane będą nosić ślady tej dyskryminacji.

Model trenowany na takich danych może osiągać wysoką trafność predykcyjną i równocześnie utrwalać strukturę, której normatywnie nie chcemy reprodukować. Historia staje się wtedy nie materiałem do uczenia się z przeszłości, lecz automatycznym projektem przyszłości. To fundamentalna różnica pomiędzy predykcją a legitymizacją. Fakt, że określona cecha pozwala przewidzieć wynik, nie oznacza, że społeczeństwo powinno zezwolić na jej wykorzystanie w procesie podejmowania decyzji. Kod pocztowy może korelować z dochodem, zdrowiem i ryzykiem kredytowym, będąc jednocześnie pośrednim nośnikiem segregacji przestrzennej. Historia zatrudnienia może przewidywać przyszłą stabilność pracy, lecz systematycznie karać osoby, których przerwy zawodowe wynikały z opieki nad dziećmi lub chorym członkiem rodziny. Matematyka identyfikuje zależność; prawo i etyka pytają, czy wolno ją instrumentalizować.

Pojęcie fairness okazuje się znacznie bardziej wymagające, niż sugeruje korporacyjne hasło o „usuwaniu biasu”. W nauce o uczeniu maszynowym istnieje wiele formalnych definicji

sprawiedliwości: równość współczynników błędów, kalibracja między grupami, równość szans, parytet selekcji czy różne odmiany niezależności predykcji od cechy chronionej. Problem polega na tym, że kryteria te mogą być ze sobą matematycznie sprzeczne. Kleinberg, Mullainathan i Raghavan wykazali, że przy realistycznych różnicach częstości bazowych nie da się jednocześnie spełnić kilku intuicyjnie atrakcyjnych własności sprawiedliwego systemu punktowego. Oznacza to coś politycznie niewygodnego: nie istnieje uniwersalny przycisk "fair". Inżynier nie może po prostu wybrać właściwej biblioteki programistycznej i ogłosić, że problem sprawiedliwości został rozwiązany. Wybór definicji fairness jest wyborem normatywnym. Czy ważniejsze jest to, aby dwa porównywalne poziomy ryzyka oznaczały to samo dla różnych grup? Czy aby odsetek fałszywie oskarżonych był podobny? A może aby grupy miały zbliżoną szansę na otrzymanie pozytywnej decyzji? Odpowiedź zależy od charakteru instytucji i moralnego znaczenia ponoszonych konsekwencji.

Algorytmiczna sprawiedliwość wymusza powrót filozofii politycznej do laboratorium informatycznego. Rawlsowska intuicja bezstronności nie podpowiada, jak ustawić wszystkie parametry klasyfikatora, lecz przypomina, że zasady decyzyjne powinny być oceniane także z perspektywy osoby, która nie wie, po której stronie rozkładu się znajdzie. Gdyby projektant nie wiedział, czy sam będzie należał do grupy częściej dotkniętej fałszywym alarmem, inaczej oceniłby społeczną dopuszczalność błędu. Z kolei Kantowska zasada traktowania człowieka jako celu, a nie wyłącznie środka, zderza się z ryzykiem redukcji osoby do profilu statystycznego: jednostka może zostać oceniona nie za to, co zrobiła, lecz za to, do jakiej probabilistycznej klasy przypomina innych. Nie oznacza to, że wszystkie decyzje statystyczne są moralnie niedopuszczalne. Ubezpieczenia, epidemiologia, medycyna czy planowanie infrastruktury od dawna opierają się na prawdopodobieństwie.

Nowość polega jednak na skali, automatyzacji i granularności. Model może oceniać miliony ludzi niemal bez kosztów krańcowych, wykorzystując tysiące cech i odkrywając relacje nieuchwytne dla ludzkiej intuicji. Tym samym zdolność klasyfikacji rośnie szybciej niż tradycyjne mechanizmy proceduralnego kwestionowania tych decyzji.

Prawo do sprawiedliwości w epoce AI musi obejmować nie tylko rezultat, lecz także możliwość zakwestionowania reprezentacji człowieka w systemie. Jeśli model ocenił mnie jako osobę wysokiego ryzyka kredytowego na podstawie błędnych danych, powinien istnieć sposób korekty. Jeśli rezultat wynika z prawidłowych danych, ale zastosowano niewłaściwe kryterium, należy zadać pytanie o legalność tego kryterium. Jeżeli zaś model działa statystycznie poprawnie, lecz jego użycie w danym kontekście jest nieproporcjonalne, problem dotyczy już samego celu.

Explainability nie powinna być redukowana do dostarczenia graficznej listy „najważniejszych cech". Wyjaśnienie techniczne i wyjaśnienie normatywne to dwie różne rzeczy. Model może wskazać, że dochód, liczba wcześniejszych zdarzeń i lokalizacja zwiększyły wynik o określone wartości, lecz obywatel ma prawo pytać: dlaczego w ogóle wolno było użyć tych cech? Dlaczego ustalono taki próg? Jaki jest współczynnik błędów w populacji podobnej do mojej? Kto ponosi odpowiedzialność za zastosowanie systemu?

Przejrzystość ma więc co najmniej trzy poziomy. Pierwszy dotyczy mechanizmu: jak system przetwarza dane. Drugi wydajności: kiedy się myli, na jakiej populacji został sprawdzony i jak zachowuje się poza środowiskiem treningowym. Trzeci dotyczy instytucji: kto zdecydował o użyciu modelu, w jakim celu, według jakich kryteriów i kto może jego decyzję zmienić. Czarna skrzynka matematyczna jest problemem, ale równie groźna bywa czarna skrzynka organizacyjna.

Współczesne podejście do bezpieczeństwa AI przesuwa uwagę z abstrakcyjnej „inteligencji" ku zarządzaniu ryzykiem. NIST AI Risk Management Framework wskazuje jako cechy godnego zaufania systemu między innymi poprawność, niezawodność, bezpieczeństwo, odporność, rozliczalność, przejrzystość, wyjaśnialność, ochronę prywatności oraz zarządzanie szkodliwym biasem. Co ważniejsze, NIST podkreśla, że cechy te należy uwzględniać w całym cyklu życia systemu – od projektowania po użycie i ewaluację. Jest to podejście bliższe inżynierii bezpieczeństwa lotniczego niż kulturze tworzenia aplikacji konsumenckich: system wysokiego ryzyka nie może zostać uznany za „dobry" tylko dlatego, że osiąga wysoki wynik na benchmarku.

Statystyczna iluzja dokładności i paradoks nadzoru ludzkiego

Pojęcie distribution shift dobrze wyjaśnia przyczynę tego zjawiska. Model uczy się zależności istniejących w konkretnym zbiorze danych, jednak po wdrożeniu w rzeczywistość może napotkać populację różniącą się pod względem wieku, zachowań, technologii, chorób, urządzeń pomiarowych czy praktyk instytucjonalnych. To, co stanowiło trafny predyktor w jednym szpitalu, mieście lub okresie, może przestać nim być w innym.

System statystyczny nie posiada metafizycznej kompetencji; posiada kompetencję warunkową względem środowiska, na którym został oceniony.

Pytanie o „dokładność modelu" jest zatem naukowo niepełne. Należy pytać: na jakiej populacji, w jakim czasie, przy jakiej częstości bazowej zjawiska, przy jakim progu, według jakiej miary, z jaką kalibracją i jaką strukturą błędów? Model osiągający 99 procent accuracy może być bezużyteczny,

jeśli zdarzenie pozytywne występuje tylko w jednym procencie przypadków, a system zawsze przewiduje jego brak. Statystyka nie pozwala więc oceniać modelu za pomocą jednej magicznej liczby.

Rozpoznawanie twarzy stanowi tutaj klasyczne laboratorium problemu. NIST wykazał znaczne różnice demograficzne między algorytmami, podkreślając, że skala tych rozbieżności zależy od konkretnego systemu, danych i zastosowania. W badaniu obejmującym 189 algorytmów wielu dostawców w części systemów występowały bardzo duże różnice w częstości fałszywych pozytywów między grupami; raport wskazywał przypadki różnic rzędu dziesięciu, a nawet ponad stu razy. Jednocześnie najlepsze algorytmy wykazywały znacznie mniej błędów, co ostrzega przed uproszczeniem, jakoby rozpoznawanie twarzy zawsze było rasistowskie. Najważniejsza lekcja jest jednak subtelniejsza.

System nie posiada jednej „dokładności rozpoznawania twarzy". Posiada rozkład błędów zależny od populacji, jakości obrazu, zadania i ustawionego progu. W systemie weryfikacji 1:1 błędne dopasowanie może dopuścić osobę nieuprawnioną. W identyfikacji 1:N fałszywy pozytyw może sprawić, że niewinny człowiek stanie się kandydatem do dalszego dochodzenia. Matematycznie ten sam błąd może mieć zupełnie inny ciężar społeczny.

Odpowiedzialna AI wymaga zatem nie tylko testowania średniej skuteczności, lecz badania heterogeniczności tej skuteczności. Jeśli średnia populacyjna maskuje grupę, wobec której system działa znacznie gorzej, „wysoka dokładność" może stać się statystycznym sposobem ukrywania niesprawiedliwości. W medycynie analogicznym błędem byłoby zachwycanie się ogólną skutecznością leczenia bez sprawdzenia, czy u określonej grupy pacjentów terapia przynosi szkody.

Z tego wynika znaczenie różnorodności zespołów projektowych, silnie podkreślane przez Smitha. Zróżnicowanie zespołu traktowane jest tu jako metoda dostrzegania „martwych pól", które jednolite środowisko może przeoczyć. Należy jednak zaznaczyć, że różnorodność nie zastępuje metodologii walidacji. Może zwiększać szansę na zadanie właściwych pytań, ale sama w sobie nie jest dowodem sprawiedliwości produktu. Konieczne są dane reprezentatywne dla celu zastosowania, audyty podgrup, analiza błędów, dokumentowanie ograniczeń oraz udział ekspertów domenowych i grup poddanych działaniu systemu.

Problem staje się jeszcze bardziej złożony w relacji AI z człowiekiem. Popularną odpowiedzią na lęk przed autonomią jest koncepcja human-in-the-loop. W źródle uznano to za jedną z najważniejszych zasad odpowiedzialności – człowiek ma pozostać w pętli decyzyjnej. Sama fizyczna obecność człowieka przy procesie nie gwarantuje jednak realnej kontroli. Psychologia automatyzacji od dawna opisuje automation bias: tendencję do nadmiernego polegania na rekomendacji systemu,

zwłaszcza gdy użytkownik uznaje go za zaawansowany, znajduje się pod presją czasu lub wykonuje zadanie trudne poznawczo. Systematyczny przegląd badań nad automation bias wykazał, że kluczowe są tu doświadczenie użytkownika, obciążenie pracą, złożoność zadania, sposób prezentacji rekomendacji oraz poziom zaufania do systemu. Nowsze badanie eksperymentalne z 2024 roku dotyczące diagnostycznego wsparcia AI wykazało, że uczestnicy potrafili zgadzać się nawet z błędnymi rekomendacjami; podatność ta była większa m.in. u osób o słabszym przygotowaniu specjalistycznym i silniejszym przekonaniu o użyteczności systemu.

Pojawia się paradoks human-in-the-loop. Im lepszy jest system, tym bardziej człowiek uczy się mu ufać. Im częściej rekomendacja okazuje się poprawna, tym trudniej zachować poznawczą czujność właśnie w rzadkich przypadkach, gdy system się myli. Nadzorca może formalnie posiadać prawo sprzeciwu, a w praktyce stać się jedynie operatorem zatwierdzającym kolejne wyniki.

W lotnictwie problem ten znany jest jako utrata kompetencji w warunkach automatyzacji. Człowiek przejmuje kontrolę najrzadziej wtedy, gdy sytuacja jest najbardziej nietypowa i wymaga kompetencji, których nie ćwiczył podczas rutynowego działania automatu. Podobne zjawisko może wystąpić w diagnostyce, analizie prawnej, administracji czy cyberbezpieczeństwie. Jeśli system przez lata wykonuje większość rozumowania wstępnego, człowiek może zachować formalne uprawnienie do interwencji, tracąc stopniowo epistemiczną zdolność jej dokonania.

Najnowsze przeglądy dotyczące AI w medycynie wskazują właśnie na tę lukę między wynikami technicznymi a klinicznym wdrożeniem. Meta-analiza opublikowana w 2026 roku, obejmująca pięćdziesiąt badań predykcyjnych systemów wsparcia klinicznego, stwierdziła znaczną heterogeniczność wyników; tylko 24 procent badań dotyczyło prospektywnego wdrożenia, a 64 procent raportowało wyłącznie metryki techniczne, pomijając informacje o rzeczywistym przebiegu pracy klinicznej.

Realna kontrola ludzka wymaga kalibracji zaufania i kompetencji poznawczych

Należy podkreślić, że model nie stanowi interwencji medycznej sam w sobie; interwencją jest model osadzony w konkretnym człowieku, procedurze, interfejsie, presji czasu i odpowiedzialności instytucjonalnej.

Realny human oversight wymaga znacznie więcej niż samego przycisku „zatwierdź”. Człowiek musi posiadać kompetencje do oceny rekomendacji, dostęp do informacji o niepewności modelu, czas na podjęcie decyzji oraz prawną i organizacyjną władzę sprzeciwu. Niezbędna jest także kultura

instytucjonalna, która nie karze za rozsądne odstąpienie od automatycznej sugestii. Jeśli pracownik wie, że każda decyzja sprzeczna z modelem będzie wymagała wielostronicowego uzasadnienia, podczas gdy zgoda przebiega automatycznie, system tworzy asymetrię zachęcającą do bezrefleksyjnego podporządkowania.

W takim układzie człowiek w pętli może stać się nie bezpiecznikiem, lecz moralnym piorunochronem dla organizacji. Decyzję faktycznie ukierunkowuje model, ale formalnie podpisuje ją człowiek, co pozwala instytucji twierdzić, że żadna decyzja nie zapadła automatycznie. To jedynie pozór odpowiedzialności. Prawdziwa kontrola wymaga zdolności zmiany wyniku, a nie tylko ceremonialnego poświadczenia.

Problem ten nabiera szczególnego znaczenia w przypadku generatywnej AI. Model językowy nie jest encyklopedią ani bazą prawdy; generuje tekst na podstawie statystycznych relacji wyuczonych z danych oraz mechanizmów dostrajania. Jego płynność językowa może być znacznie wyższa niż pewność epistemiczna. Powstaje niebezpieczna asymetria: człowiek naturalnie interpretuje pewny, uporządkowany i profesjonalny styl jako sygnał kompetencji, podczas gdy mechanizm generowania nie gwarantuje zgodności tej pewności stylistycznej z prawdopodobieństwem prawdziwości. Prowadzi to do nowej kategorii ryzyka: błędu perswazyjnego.

Tradycyjny program komputerowy zwracał kod błędu lub błędną liczbę. Model generatywny może natomiast podać fałszywą odpowiedź w formie świetnie skonstruowanego uzasadnienia. Błąd otrzymuje retorykę. W administracji, prawie, medycynie czy nauce jest to szczególnie niebezpieczne, gdyż człowiek może zaufać odpowiedzi nie dlatego, że została lepiej zweryfikowana, lecz dlatego, że została lepiej napisana.

W tym miejscu epistemologia spotyka się z psychologią poznawczą. Człowiek rzadko ocenia każdą informację od zera; posługuje się heurystykami. Profesjonalny ton, szybkość odpowiedzi, szczegółowość i spójność narracji działają jako sygnały wiarygodności, co sprawia, że AI może nieumyślnie wykorzystywać nasze poznawcze skróty.

Odpowiedzialne projektowanie powinno zatem nie tylko zwiększać jakość odpowiedzi, ale kalibrować zaufanie człowieka do tej jakości. Najlepszy system nie jest tym, któremu ufa się zawsze, lecz tym, któremu ufa się właściwie: bardziej wtedy, gdy system ma mocne podstawy, i mniej, gdy działa poza obszarem swoich kompetencji. Celem nie powinno być maksymalne trust, lecz calibrated trust. Nadmierna nieufność marnuje korzyści z automatyzacji, nadmierne zaufanie zaś przekształca narzędzie wsparcia w niewidzialnego decydenta.

Najostrzejszy wariant tego problemu pojawia się w domenie wojskowej. Debata wokół Google i Project Maven oraz decyzja Microsoftu o współpracy z sektorem obronnym wskazują na

„meaningful human control” jako granicę szczególnej wagi. Przeciwstawienie „pacyfizmu Google” i „wojny sprawiedliwej Microsoftu” warto jednak traktować jako retoryczne uproszczenie, a nie naukową klasyfikację.

Historycznie Project Maven koncentrował się na wykorzystaniu computer vision do automatycznego wykrywania i klasyfikowania obiektów w ogromnych zasobach obrazu i wideo, aby odciążyć analityków. Pentagon przedstawiał go jako projekt percepcyjny wspierający człowieka, a nie autonomiczną broń wybierającą cel i otwierającą ogień. Google w 2018 roku zdecydował, że nie będzie ubiegać się o dalszy kontrakt Maven, publikując jednocześnie własne zasady AI. Microsoft natomiast wszedł później we współpracę z US Army przy IVAS – systemie zwiększającym świadomość sytuacyjną żołnierza poprzez łączenie sensorów, rozszerzonej rzeczywistości i uczenia maszynowego.

Spór nie dotyczy więc jedynie pytania o dopuszczalność technologii w wojsku, lecz tego, jak daleko w łańcuchu użycia siły może sięgać automatyzacja, zanim ludzki sąd stanie się fikcją. Można wyróżnić tu kilka etapów: system może wykryć obiekt, sklasyfikować go, oszacować zagrożenie, zaproponować priorytet, rekomendować reakcję, wybrać cel i wreszcie zastosować siłę. Przejście między tymi etapami nie jest neutralne; każdy kolejny krok przesuwa system od automatyzacji percepcji ku automatyzacji intencji. Największy problem moralny pojawia się wtedy, gdy probabilistyczne rozpoznanie zostaje bez wystarczającej ludzkiej refleksji przekształcone w nieodwracalne działanie. W wojnie czas jest jednak zasobem strategicznym – przeciwnik wykorzystujący AI może szybciej identyfikować zagrożenia, integrować sensory i proponować reakcję.

Iluzja kontroli i problem rozproszonej odpowiedzialności

Powstaje zatem presja na "kompresję pętli decyzyjnej". Im szybciej maszyna przetwarza sytuację, tym bardziej człowiek wydaje się wąskim gardłem. To właśnie w tym punkcie techniczna przewaga wchodzi w konflikt z etycznym wymogiem namysłu; maszyna skraca czas reakcji dokładnie tam, gdzie moralność potrzebuje przestrzeni na wątpliwość. Z tego powodu "meaningful human control" nie może sprowadzać się do samej obecności człowieka w strukturze dowodzenia. Sensowna kontrola wymaga, aby operator rozumiał naturę systemu, dysponował pełnym kontekstem, potrafił przewidzieć konsekwencje, posiadał realną możliwość interwencji i ponosił odpowiedzialność za użycie siły. ICRC od lat podkreśla, że nieprzewidywalność systemów autonomicznych rośnie wraz

ze złożonością środowiska oraz ograniczaniem ludzkiego nadzoru, a uczenie maszynowe dodatkowo utrudnia wyjaśnienie zachowania systemu.

Debata ta przestała być domeną abstrakcyjnej futurystyki. W lipcu 2026 roku António Guterres ponowił postulat, by systemy zdolne do wybierania i atakowania ludzi bez ludzkiego osądu zostały zakazane prawem międzynarodowym; podkreślił wprost, że decyzje o odebraniu życia muszą pozostać ludzkie. W tym samym roku prace w ramach Convention on Certain Conventional Weapons nadal koncentrują się na kwestii kontroli nad użyciem siły. Problem filozoficzny jest jednak głębszy niż sama skuteczność techniczna.

Człowiek odpowiedzialny za użycie siły może zostać przesłuchany, osądzony i zmuszony do uzasadnienia swojej decyzji w świetle norm prawnych. Maszyna nie jest podmiotem moralnym. Nie doświadcza winy, nie rozumie godności przeciwnika i nie może ponieść kary przed sądem. Dlatego nawet techniczna perfekcja nie rozwiązuje kwestii odpowiedzialności – nie można delegować obowiązku moralnego na podmiot, który z definicji nie może go posiadać. Zasada ta znajduje zastosowanie również w sferze cywilnej, choć stawka jest inna.

Bank nie może zasłaniać się tym, że "algorytm odmówił kredytu", szpital że "AI postawiła diagnozę", a urząd że "system nie przyznał świadczenia", skoro to organizacja zdecydowała o wdrożeniu narzędzia i jego parametrach. Model nie jest nowym urzędnikiem stojącym poza strukturą odpowiedzialności, lecz częścią instrumentarium instytucji. Tutaj ujawnia się socjologiczny problem many hands – zjawisko odpowiedzialności rozproszonej. Dane zebrał jeden podmiot, model stworzył drugi, dostroił trzeci, zintegrował czwarty, wdrożyła organizacja piąta, a decyzję formalnie podpisał pracownik szósty. W takim układzie każdy może wskazać na kogoś innego.

Inżynier twierdzi, że nie znał konkretnego zastosowania; dostawca, że klient błędnie ustawił parametry; klient, że zaufał renomowanemu modelowi; pracownik zaś, że postępował zgodnie z rekomendacją systemu. Jeśli odpowiedzialność zostanie rozdzielona bez zaprojektowania jej architektury, może paradoksalnie zniknąć. Sto osób odpowiedzialnych częściowo może stworzyć system, za który nie odpowiada nikt w całości.

Przejście od etyki dobrej woli do twardych ram prawnych

Dlatego governance AI wymaga nie tylko zasad etycznych, lecz precyzyjnej mapy kompetencji: określenia, kto odpowiada za dane, walidację, dobór celu, bezpieczeństwo, monitorowanie driftu,

decyzję wdrożeniową, ludzką kontrolę oraz procedurę wycofania systemu. To również powód, dla którego postulat "przysięgi Hipokratesa dla programistów", obecny w książce Smitha, jest intelektualnie interesujący, ale niewystarczający. Medycyna nie opiera bezpieczeństwa wyłącznie na moralności lekarza; posiada procedury licencyjne, badania kliniczne, standardy zawodowe, komisje etyczne, system odpowiedzialności, nadzór i obowiązki dokumentacyjne.

Jeżeli inżynieria AI rzeczywiście staje się profesją zdolną wpływać na życie i prawa milionów ludzi, odpowiedniki takich instytucji są ważniejsze niż sama deklaracja dobrej woli. Należy przy tym wystrzegać się przeciwnego błędu: przekonania, że AI jest z natury mniej moralna niż człowiek. Ludzkie decyzje również bywają stronnicze, niespójne, impulsywne i zależne od zmęczenia.

Lekarze, sędziowie, rekruterzy czy urzędnicy nie są bezbłędnymi punktami odniesienia. Algorytm może ujawnić ludzką niespójność, a w niektórych zastosowaniach nawet ją zredukować. Dlatego właściwym pytaniem nie jest "człowiek czy maszyna?", lecz jaka konfiguracja człowieka, danych, modelu i procedury zapewnia lepszy wynik przy akceptowalnej strukturze ryzyka. Zmienia to sens human-centred AI. Człowiek nie musi wykonywać każdej czynności ręcznie, aby system był humanistyczny. Humanizm technologiczny polega raczej na tym, że człowiek pozostaje źródłem wartości, dla których system działa, oraz adresatem praw chroniących go przed błędem systemu. Automatyzacja może być bardzo głęboka, o ile granice jej zastosowania pozostają społecznie kontrolowane. Najdojrzalsza formuła nie brzmi więc "AI ma zawsze słuchać człowieka", ponieważ człowiek także może wydawać niemoralne polecenia. Brzmi raczej: AI i człowiek muszą pozostawać wewnątrz systemu norm, które żadnemu z nich nie pozwalają na niekontrolowaną władzę. To samo państwo prawa, które ogranicza urzędnika, musi nauczyć się ograniczać system wspomagający urzędnika.

W tym miejscu można już dostrzec zmianę epoki. Pierwsza generacja debaty o odpowiedzialnej AI miała charakter przede wszystkim etyczny: przedsiębiorstwa publikowały zasady fairness, transparency, accountability, privacy czy safety. Książka Smitha dobrze dokumentuje ten moment poszukiwania swoistej "konstytucji" dla technologii rozwijającej się szybciej niż państwowe regulacje. Jednak historia prawa pokazuje, że normy społeczne zaczynają jako zalecenia, później stają się standardami zawodowymi, a ostatecznie część z nich przechodzi w obowiązki prawne.

Ten proces obserwujemy dziś. Fairness przestaje być wyłącznie deklaracją etyczną i staje się przedmiotem obowiązków dotyczących zarządzania ryzykiem. Transparency ewoluuje od manifestu do obowiązku informowania. Human oversight, będący dotąd postulatem filozoficznym, przechodzi w konstrukcję regulacyjną. Dokumentowanie danych, testowanie, monitoring i odpowiedzialność zaczynają podlegać prawu publicznemu. Jest to moment, w którym społeczeństwo mówi technologii coś bardzo starego: możliwość działania nie jest jeszcze uprawnieniem do działania. Po

analizie filozofii sprawiedliwości, statystyki błędu, psychologii automatyzacji, medycyny, biometrii i etyki wojny trzeba zapytać, co z tych zasad zostało już przełożone z języka dobrowolnej odpowiedzialności na język normy obowiązującej. Etyka staje się prawem. AI Act i narodziny publicznej odpowiedzialności za sztuczną inteligencję stanowią odpowiedź na ten proces.

Dotarliśmy do granicy, na której etyczny postulat przestaje wystarczać. Jeżeli system sztucznej inteligencji może wpływać na dostęp człowieka do pracy, edukacji, kredytu, świadczeń, granicy państwowej, informacji, opieki medycznej albo wymiaru sprawiedliwości, pytanie o odpowiedzialność nie może być pozostawione wyłącznie dobrej woli producenta.

Przejście od etyki dobrej woli do twardych regulacji prawnych

Historia techniki wielokrotnie przebiegała według podobnego schematu. Najpierw pojawiało się nowe narzędzie, potem pierwsze szkody i kodeksy dobrych praktyk, następnie normy zawodowe, standardy techniczne oraz kontrola instytucjonalna, by na końcu sformułować obowiązek prawny. Mosty, samoloty, leki czy elektrownie nie są bezpieczne dlatego, że ich konstruktorzy podpisali deklarację etyczną. Bezpieczeństwo staje się tam procedurą: dokumentacją, testem, audytem, kompetencją organu nadzoru i realną możliwością nałożenia sankcji.

Dokładnie ten sam proces zaczyna obejmować sztuczną inteligencję. Książka Smitha dokumentuje wcześniejszy etap tej ewolucji: próbę zbudowania korporacyjnej „konstytucji AI" w oparciu o zasady sprawiedliwości, niezawodności, bezpieczeństwa, prywatności, inkluzywności, przejrzystości oraz odpowiedzialności, przypisując szczególną rolę zachowaniu człowieka w pętli decyzyjnej. W roku 2026 znaczna część tego słownika przestała należeć wyłącznie do etyki przedsiębiorstw i została przejęta przez prawodawcę. Jest to przejście od soft law do hard law, którego nie należy jednak rozumieć jako prostego zastąpienia moralności kodeksem. Etyka pyta o to, co powinno być czynione nawet wtedy, gdy nie grozi kara. Prawo wyznacza minimalny poziom zachowania, którego społeczeństwo może przymusowo wymagać. Standard techniczny z kolei próbuje przełożyć abstrakcyjne wymaganie na powtarzalną praktykę inżynieryjną. W dojrzałym systemie te trzy warstwy powinny się uzupełniać: etyka określa horyzont, prawo granicę dopuszczalności, a standaryzacja metodę wykazania, że granica ta została zachowana. Unijny AI Act jest szczególnie interesujący właśnie dlatego, że nie próbuje uregulować „sztucznej inteligencji" jako jednolitego przedmiotu, lecz przyjmuje architekturę opartą na ryzyku.

Oznacza to przejście od ontologicznego pytania „czy to jest AI?” do funkcjonalnego zapytania: „co może się wydarzyć, gdy ten system zostanie użyty w określonym kontekście?”. Jest to podejście intelektualnie dojrzalsze niż traktowanie każdego algorytmu w ten sam sposób. Filtr antyspamowy i system wspierający decyzję o pozbawieniu człowieka wolności należą do tej samej rodziny technologicznej, lecz nie powinny podlegać jednakowej intensywności kontroli. Przypomina to regulacje bezpieczeństwa w medycynie i technice: nie każdy termometr wymaga procedur właściwych dla rozrusznika serca, a nie każda śruba lotnicza jest oceniana tak samo jak cały system sterowania samolotem. Im poważniejszy potencjalny skutek, tym większa intensywność kontroli ex ante. AI Act przenosi tę intuicję do przestrzeni informacji i decyzji.

Tutaj pojawia się jednak pierwsza różnica między bezpieczeństwem klasycznego produktu a bezpieczeństwem AI. Uszkodzony hamulec oddziałuje przede wszystkim fizycznie; algorytm może natomiast oddziaływać epistemicznie i normatywnie: zaklasyfikować, stworzyć reputację, ograniczyć widzialność, wzmocnić podejrzenie czy przesunąć człowieka do kategorii ryzyka. Szkoda może zatem powstać bez jakiegokolwiek uszkodzenia materialnego. AI Act musi chronić nie tylko zdrowie i bezpieczeństwo w sensie technicznym, lecz także prawa podstawowe.

Europejska regulacja AI stanowi interesujące połączenie dwóch tradycji prawnych. Pierwsza wywodzi się z europejskiego prawa bezpieczeństwa produktów: analiza ryzyka, ocena zgodności, dokumentacja, normy techniczne i nadzór rynku. Druga pochodzi z konstytucjonalizmu praw człowieka: godność, niedyskryminacja, prywatność, sprawiedliwa procedura, autonomia oraz ochrona przed arbitralnością. W AI Act inżynieria produktu spotyka się z filozofią praw podstawowych.

Radykalną techniką regulacji jest oczywiście zakaz. Unia uznała, że pewne zastosowania AI generują ryzyko na tyle niezgodne z porządkiem praw podstawowych, że nie powinno się ich jedynie „łagodzić”. Od 2 lutego 2025 roku stosowane są zakazy obejmujące między innymi określone formy szkodliwej manipulacji i wykorzystywania podatności człowieka, social scoring, indywidualne przewidywanie ryzyka popełnienia przestępstwa oparte wyłącznie na profilowaniu, niektóre formy rozpoznawania emocji w miejscu pracy i edukacji, szczególnie wrażliwą kategoryzację biometryczną oraz pewne zastosowania zdalnej identyfikacji biometrycznej.

Filozofia tego rozwiązania jest znacząca. Prawodawca nie postuluje budowy „bardziej dokładnego social scoringu”, lecz stwierdza, że w określonych konfiguracjach sama idea takiego systemu narusza granicę dopuszczalnej władzy nad człowiekiem. Oznacza to powrót kategorii nienaruszalnego tabu prawnego do świata technologii. Nie wszystko, co da się zoptymalizować, powinno być optymalizowane. Jest to odpowiedź na pułapkę technokratycznej racjonalności, w której technologia ma naturalną tendencję do przekształcania problemu normatywnego w problem

wydajności. Jeżeli system dyskryminuje, inżynier może zapytać, jak poprawić model. Prawo musi jednak czasem zadać pytanie wcześniejsze: czy w ogóle wolno budować ten rodzaj klasyfikacji? Właśnie tutaj społeczeństwo odzyskuje prymat nad techniką. Nie każda wadliwa technologia potrzebuje lepszego modelu; niektóre zastosowania po prostu wymagają zakazu.

Przekład etyki na twarde obowiązki operacyjne i dowodowe

Drugą warstwę stanowią systemy wysokiego ryzyka. Z perspektywy całego wywodu jest to konstrukcja kluczowa, gdyż obrazuje proces przekładania pojęć etycznych na konkretne obowiązki operacyjne. W docelowym reżimie dostawca takiego systemu będzie zobligowany do wdrożenia systemu zarządzania ryzykiem, zapewnienia wysokiej jakości danych, prowadzenia dokumentacji technicznej i logowania oraz dostarczenia niezbędnych informacji podmiotowi wdrażającemu. Musi on również zagwarantować adekwatny nadzór człowieka a także wymagany poziom dokładności, odporności i cyberbezpieczeństwa. Warto przeanalizować naturę tej transformacji.

Fairness ewoluuje w wymogi dotyczące danych i mitygacji ryzyka dyskryminacji. Accountability materializuje się jako dokumentacja, rejestrowanie zdarzeń i możliwość audytu. Human oversight przestaje być jedynie hasłem, a staje się elementem konstrukcyjnym systemu. Reliability and safety przekształcają się w twarde wymagania techniczne. Transparency z kolei przestaje oznaczać dobrą praktykę komunikacyjną, a staje się dostarczeniem informacji niezbędnych do prawidłowego użycia i kontroli narzędzia. Prawo dokonuje tu czegoś, czego sama etyka nie potrafi: tworzy ślady odpowiedzialności. System wysokiego ryzyka nie może być po prostu „dobry”.

Konieczne jest, aby można było jednoznacznie wykazać, kto go stworzył, w jakim celu, na jakich założeniach, w jaki sposób został zweryfikowany, jakie posiada ograniczenia i jakie zdarzenia wystąpiły podczas jego pracy. W społeczeństwie zautomatyzowanym logowanie systemowe może stać się odpowiednikiem akt postępowania administracyjnego. Bez śladu decyzji nie ma bowiem skutecznego prawa do jej kontroli. Ta analogia jest głębsza, niż mogłoby się wydawać: państwo prawa od dawna wymaga, aby organ nie tylko podjął decyzję, ale był zdolny wykazać podstawę kompetencji, zgromadzony materiał oraz tok rozumowania, umożliwiając tym samym zaskarżenie aktu. System AI uczestniczący w procesie decyzyjnym nie może tworzyć epistemicznej dziury wewnątrz tej procedury. Jeśli obywatel otrzymuje decyzję administracyjną podpisaną przez urzędnika, lecz jej rzeczywistym motorem była niemożliwa do odtworzenia klasyfikacja algorytmiczna, formalna procedura zachowuje kształt, tracąc jednak swoją treść.

Logowanie i dokumentacja nie są zatem technokratycznym obciążeniem. Stanowią warunek rekonstrukcji odpowiedzialności. Prawo bez dowodu jest bezsilne; odpowiedzialność pozbawiona możliwości odtworzenia działania systemu staje się fikcją.

Istotnym elementem europejskiego modelu jest również ocena zgodności. Zanim system wysokiego ryzyka zostanie wprowadzony do obrotu lub oddany do użytku, dostawca musi wykazać jego zgodność z wymogami. Przy istotnych zmianach systemu lub celu ocena może wymagać powtórzenia, a w zależności od kategorii systemu może w proces ten zaangażowana zostać niezależna jednostka oceniająca. Oznacza to fundamentalną zmianę kultury wytwarzania oprogramowania. Przez dekady sektor cyfrowy rozwijał się według logiki ship, measure, iterate: wypuść produkt, obserwuj użytkowników, szybko poprawiaj.

Logika prewencji i infrastruktura wykonawcza prawa

Metoda ta sprawdza się w wielu aplikacjach niskiego ryzyka. Trudno ją jednak zaakceptować tam, gdzie błędem eksperymentu jest naruszenie czyjegoś prawa, utrata świadczenia lub błędna ocena medyczna. Prawo wysokiego ryzyka mówi w istocie: nie każde społeczne wdrożenie może być testem A/B wykonywanym na obywatelach. Różnica pomiędzy oprogramowaniem konsumenckim a systemem regulowanym dotyczy zatem momentu przejęcia odpowiedzialności. Kultura start-upowa często przesuwa ją na etap po wdrożeniu: najpierw innowacja, później reakcja na problem. Regulacja wysokiego ryzyka wymaga natomiast odpowiedzialności przed wprowadzeniem systemu. To logika prewencji.

Nie jest jednak prawdą, że od 2 sierpnia 2026 roku wszystkie wymagania dotyczące systemów wysokiego ryzyka zaczęły już działać w pełnym zakresie. Stan prawny roku 2026 wymaga szczególnej precyzji. AI Act jako całość wszedł zasadniczo w fazę stosowania 2 sierpnia 2026 roku, lecz po wejściu w życie AI Omnibus 27 lipca 2026 roku terminy dla reżimu wysokiego ryzyka zostały przesunięte. Wymagania dotyczące zastosowań wymienionych w załączniku III – między innymi w biometrii, edukacji, zatrudnieniu, migracji, egzekwowaniu prawa czy wymiarze sprawiedliwości – mają być stosowane od 2 grudnia 2027 roku, natomiast dla AI będącej częścią produktów objętych załącznikiem I termin ten sięga 2 sierpnia 2028 roku.

Ta zmiana jest pouczająca z punktu widzenia teorii regulacji. Nie oznacza ona rezygnacji z modelu wysokiego ryzyka, lecz pokazuje, że prawo samo potrzebuje infrastruktury, aby stać się wykonalne. Niezbędne są standardy zharmonizowane, wytyczne, jednostki oceny zgodności, kompetentne organy nadzoru, procedury i praktyka audytowa. Ustawa może zostać uchwalona jednym głosowaniem, ale system zgodności musi zostać zbudowany. To znany paradoks regulacji

technologicznej: jeśli prawodawca działa zbyt wolno, normuje świat, który już nie istnieje; jeśli działa zbyt szybko, może wprowadzić obowiązki, dla których rynek, administracja i system standardów nie posiadają jeszcze narzędzi wykonawczych. AI Omnibus jest próbą korekty tego napięcia. Komisja przedstawia wydłużenie terminów jako element uproszczenia i zwiększania pewności prawnej, równolegle rozszerzając sandboksy regulacyjne i mechanizmy wsparcia dla mniejszych podmiotów.

Czas regulacji staje się zatem elementem samej polityki innowacyjnej. Zbyt wczesne wprowadzenie skomplikowanych wymagań może uprzywilejować największe przedsiębiorstwa, dysponujące rozbudowanymi działami prawnymi, compliance i bezpieczeństwa. Mały podmiot może posiadać znakomitą technologię, lecz nie być w stanie ponieść stałego kosztu wielowarstwowej dokumentacji i oceny zgodności. Regulacja zaprojektowana dla ochrony społeczeństwa może niezamierzenie zwiększyć koncentrację rynku.

Z drugiej strony brak wspólnych standardów również działa na korzyść największych. Mały nabywca technologii nie potrafi samodzielnie audytować wielkiego modelu. Szpital, szkoła czy samorząd nie mają zasobów potrzebnych do tworzenia całej metodologii oceny bezpieczeństwa AI. Standaryzacja może obniżyć koszt transakcyjny zaufania: zamiast każdego klienta prowadzącego własne quasi-dochodzenie otrzymujemy wspólną procedurę wykronywania zgodności.

Zatem odpowiedź na pytanie "czy regulacja hamuje innowację?" nie może być binarna. Dobra regulacja może hamować niektóre zachowania, a jednocześnie umożliwiać rozwój całego rynku poprzez zwiększanie przewidywalności i zaufania. Normy bezpieczeństwa lotniczego niewątpliwie podnoszą koszt skonstruowania samolotu, ale świat bez wspólnych standardów nie stworzyłby bardziej dynamicznego lotnictwa pasażerskiego – prawdopodobnie stworzyłby rynek mniejszy, bardziej chaotyczny i mniej godny zaufania. W tej perspektywie harmonizowane standardy CEN i CENELEC są jednym z najmniej efektownych medialnie, a zarazem kluczowych elementów europejskiego systemu AI. Mają one obejmować między innymi zarządzanie ryzykiem, jakość danych, prowadzenie zapisów, transparentność, nadzór człowieka, dokładność, odporność, cyberbezpieczeństwo, system zarządzania jakością i ocenę zgodności.

Standardy techniczne i regulacja modeli GPAI jako fundament kontroli

Stosowanie zharmonizowanego standardu ma być dobrowolne, lecz może tworzyć domniemanie zgodności z odpowiednim wymaganiem prawnym. Powstaje w ten sposób ciekawa forma

delegowanego konstytucjonalizmu technicznego. Parlament określa wartość: system ma być bezpieczny i podlegać nadzorowi człowieka. Eksperci normalizacyjni muszą następnie przełożyć tę wartość na praktykę inżynierską. Jaki test jest wystarczający? Jak dokumentować dane? Jak mierzyć odporność? I wreszcie – jak udowodnić, że człowiek rzeczywiście może przejąć kontrolę?

Standard techniczny staje się wówczas niewidzialnym miejscem translacji między polityką a kodem. Jeśli zostanie źle zaprojektowany, może formalnie realizować ustawę, a w rzeczywistości wypaczać jej cel. Dlatego demokratyczna kontrola nad AI nie kończy się na przyjęciu rozporządzenia. Znaczna część realnej normy kształtuje się później: w laboratoriach, komitetach standaryzacyjnych, wytycznych organów i praktyce audytowej.

Inny problem ujawnił rozwój modeli ogólnego przeznaczenia. Wczesna logika regulacji AI zakładała stosunkowo prosty układ: znany system, konkretny dostawca i określone zastosowanie. Generatywna AI rozbiła ten model. Ten sam system może stać się asystentem prawnika, generatorem kodu, nauczycielem, komponentem systemu medycznego lub narzędziem marketingowym. Producent modelu nie zawsze kontroluje jego końcowe użycie, z kolei dostawca downstream nie posiada pełnej wiedzy o procesie treningu i ograniczeniach architektury. Powstaje zatem łańcuch odpowiedzialności podobny do przemysłowych łańcuchów dostaw, lecz przedmiotem dostawy jest zdolność poznawcza.

Z tego powodu AI Act wprowadził odrębny reżim dla modeli general-purpose AI (GPAI). Od 2 sierpnia 2025 roku ich dostawcy podlegają m.in. obowiązkom dokumentacyjnym, wymogowi przekazywania informacji podmiotom downstream, konieczności posiadania polityki zgodności z unijnym prawem autorskim oraz publikowania szczegółowego podsumowania treści treningowych. Modele uznane za niosące ryzyko systemowe podlegają dodatkowo rygorom oceny i ograniczania ryzyka, raportowania incydentów oraz cyberbezpieczeństwa. Od 2 sierpnia 2026 roku Komisja będzie mogła egzekwować pełne przestrzeganie tych obowiązków przez dostawców nowych modeli GPAI, w tym nakładać sankcje.

Jest to moment historycznie istotny: regulacja nie czeka już wyłącznie na pojawienie się konkretnego, szkodliwego produktu downstream. Zaczyna kontrolować warstwę podstawową, na której opierać się będą liczne systemy. Można to porównać do regulacji materiałów konstrukcyjnych. Producent betonu nie zna każdego przyszłego budynku, ale musi dostarczyć precyzyjne parametry materiału, gdyż bez nich architekt nie jest w stanie odpowiedzialnie zaprojektować konstrukcji. Podobnie dostawca modelu ogólnego przeznaczenia nie zna każdej późniejszej implementacji, lecz nie może zasłonić się tym faktem, by uniknąć obowiązku przekazania informacji o możliwościach, ograniczeniach i charakterze systemu.

Transparentność pochodzenia jako infrastruktura proweniencji

Współodpowiedzialność łańcuchowa stanowi istotne antidotum na problem "many hands", opisany wcześniej. W sytuacji, gdy każda warstwa ekosystemu może zrzucić odpowiedzialność na inną, w efekcie nie odpowiada nikt. AI Act próbuje temu przeciwdziałać poprzez wprowadzenie odpowiedzialności warstwowej: osobne obowiązki spoczywają na twórcy modelu podstawowego, integratorze, dostawcy systemu oraz deployerze.

Jeszcze bardziej bezpośrednio odczuwalna staje się nowa warstwa przejrzystości interakcji i treści syntetycznych, która weszła w życie 2 sierpnia 2026 roku. To właśnie tego dnia zaczął być stosowany artykuł 50 AI Act. Zgodnie z nim, interaktywne systemy AI muszą w odpowiednich przypadkach informować człowieka o tym, że komunikuje się on z maszyną. Deepfake'i podlegają obowiązkowemu oznaczaniu, a dostawcy określonych systemów generatywnych są zobligowani do umożliwienia wykrywania treści syntetycznych lub zmodyfikowanych poprzez oznaczenia maszynowo odczytywalne.

Na pierwszy rzut oka jest to obowiązek skromny. Nie zakazuje on generowania treści, nie rozstrzyga o prawdziwości przekazu ani nie ustanawia urzędu kontroli informacji. Wymaga jednak czegoś epistemologicznie podstawowego: oznacz źródło ontologiczne komunikatu. Ma to fundamentalne znaczenie w środowisku, w którym dotychczasowa heurystyka "widzę fotografię, więc wydarzenie musiało wyglądać mniej więcej tak" traci swoją wiarygodność. Przez ponad sto lat społeczeństwo wypracowało bowiem szczególną kulturę dowodową opartą na fotografii i nagraniu.

Generatywna AI drastycznie zwiększa koszt korzystania z tej heurystyki. Nie likwiduje ona dowodu wizualnego, lecz wymusza wprowadzenie nowych mechanizmów ustalania jego pochodzenia. Obowiązek oznaczania syntetyczności jest zatem analogiczny do wcześniejszych technologii weryfikacji dokumentu: podpisu, pieczęci, nagłówka wydawcy czy podpisu cyfrowego. Nie odpowiada on na pytanie, czy treść jest prawdziwa, lecz pomaga określić, jakiego rodzaju obiekt poznawczy znajduje się przed odbiorcą.

Oznaczenie maszynowo odczytywalne (machine-readable marking) jest w tym kontekście ważniejsze niż wizualny napis "AI-generated". Znak widzialny dla człowieka może zostać bowiem usunięty podczas kopiowania, kadrowania lub ponownej publikacji. Oznaczenie wykrywalne maszynowo ma natomiast pozwolić ekosystemowi informacyjnemu na automatyczne rozpoznawanie pochodzenia treści. Prawo zaczyna więc wymagać nie tylko przejrzystości semantycznej, ale infrastruktury proweniencji. Tutaj jednak ustawodawca musi unikać budowania

fałszywego poczucia bezpieczeństwa. Oznaczenie "AI-generated" nie jest tożsame z etykietą "fałszywe", podobnie jak jego brak nie dowodzi autentyczności. Człowiek może skłamać bez użycia AI, a model może wygenerować perfekcyjnie prawdziwy opis. Transparentność pochodzenia jest zatem tylko jednym z sygnałów epistemicznych, a nie maszyną do rozstrzygania prawdy.

Aktualne wytyczne Komisji z lipca 2026 roku doprecyzowują te obowiązki, obejmując m.in. chatboty, deepfake'i, określone treści dotyczące spraw interesu publicznego oraz zastosowania w zakresie rozpoznawania emocji i kategoryzacji biometrycznej. Dla części systemów wprowadzonych na rynek przed 2 sierpnia 2026 roku przewidziano ograniczony okres przejściowy dotyczący technicznego oznaczania treści, który trwa do 2 grudnia 2026 roku. Widać zatem, że rok 2026 nie jest pojedynczym momentem wejścia w życie AI Act.

Jest to raczej sekwencja warstw regulacyjnych. Zakazane praktyki zaczęły obowiązywać najwcześniej, od lutego 2025 roku. Obowiązki dotyczące GPAI weszły w fazę stosowania w sierpniu 2025 roku, a pełne egzekwowanie wobec nowych modeli rozpoczęło się w sierpniu 2026. Od 2 sierpnia 2026 działają także zasadnicze obowiązki transparentności z art. 50. Regulacje dotyczące systemów wysokiego ryzyka zostały natomiast przesunięte na lata 2027-2028. To wieloetapowe wdrożenie jest istotne również z perspektywy socjologii prawa: norma nie pojawia się bowiem w przestrzeni społecznej w dniu publikacji Dziennika Urzędowego.

Od etyki dobrej woli do twardego nadzoru epistemicznego

Organizacje muszą zinterpretować te przepisy, powołać zespoły compliance oraz zrewidować dokumentację, architekturę produktów, procedury zakupowe i strukturę odpowiedzialności zarządczej. Regulator musi natomiast wypracować kompetencje pozwalające kontrolować technologię znacznie bardziej złożoną niż systemy, które nadzorował dotychczas. Powstaje asymetria wiedzy charakterystyczna dla regulacji obszarów eksperckich: podczas gdy regulator dysponuje kompetencją prawną, przedsiębiorstwo posiada głębszą wiedzę techniczną o samym systemie. Dlatego nadzór nad AI wymaga synergii prawa, inżynierii, statystyki, cyberbezpieczeństwa, ekonomii i nauk społecznych. Prawnik pozbawiony wiedzy o modelach może nie rozpoznać rzeczywistego ryzyka; z kolei inżynier bez przygotowania prawnego może przeoczyć fakt, że zmiana jednego parametru systemu zmienia charakter ingerencji w prawa człowieka.

Urząd Komisji Europejskiej ds. AI oraz organy krajowe stają się tym samym nowym rodzajem administracji epistemicznej. Ich zadaniem nie będzie jedynie weryfikacja poprawności złożonych

formularzy, lecz ocena systemów probabilistycznych, modeli ogólnego przeznaczenia, dokumentacji treningowej, strategii mitygacji ryzyka oraz relacji między technologią a prawami podstawowymi. Od 2 sierpnia 2026 roku Komisja i organy krajowe weszły w nową fazę egzekwowania AI Act, przy czym AI Office posiada szczególne kompetencje wobec GPAI i wybranych systemów. Wprowadzenie sankcji zmienia ekonomię odpowiedzialności. Naruszenia zakazanych praktyk mogą skutkować karami do 35 mln euro lub 7% światowego rocznego obrotu, natomiast inne uchybienia – do 15 mln euro lub 3% obrotu, zależnie od kategorii i konkretnych przepisów. Prawo przestaje zatem jedynie prosić o responsible AI; w określonym zakresie zaczyna ją wymuszać. Nie należy jednak mylić surowości sankcji ze skutecznością regulacji – kara jest bowiem ostatnim ogniwem procesu.

Jeśli definicje pozostaną niejasne, standardy niedojrzałe, a audyty powierzchowne i organy nieposiadające odpowiednich kompetencji będą je nadzorować, wysoki wymiar kar może mieć znaczenie głównie symboliczne. Skuteczne prawo technologiczne wymaga od regulatora zdolności poznawczej, a nie tylko aparatu represyjnego.

Europejski model należy oceniać także krytycznie. Klasyfikacja ryzyka jest zawsze uproszczeniem: system uznany za niskiego ryzyka może w dużej skali generować szkody społeczne, podczas gdy rozwiązanie wysokiego ryzyka, stosowane odpowiedzialnie przez profesjonalną instytucję, może przynosić ogromne korzyści. Kategorie prawne muszą być na tyle stabilne, by przedsiębiorcy znali swoje obowiązki, a jednocześnie na tyle elastyczne, by nie zamrozić stanu technologii z roku uchwalenia ustawy w normę obowiązującą przez dekadę. Pojawia się również problem ryzyka emergentnego. Model ogólnego przeznaczenia może po aktualizacji ujawnić zdolności, których wcześniej nie przewidywano; sposób jego użycia może ewoluować bez formalnej zmiany produktu, a połączenie z innym systemem może stworzyć zupełnie nową funkcjonalność.

Tradycyjne prawo bezpieczeństwa produktów zakłada stabilność przedmiotu obrotu. AI jest natomiast często usługą podlegającą ciągłym modyfikacjom. Przyszłość regulacji będzie zatem wymagała przejścia od jednorazowej certyfikacji ku ciągłemu nadzorowi cyklu życia. Pytanie o zgodność systemu w momencie wejścia na rynek stanie się niewystarczające. Kluczowe będzie badanie, czy system pozostaje zgodny po aktualizacji, zmianie danych, integracji z innymi usługami czy zmianie populacji użytkowników. Oznacza to również redefinicję prawa do innowacji. Innowacja nie może być utożsamiana wyłącznie z szybkością wdrażania nowych funkcji. W systemie dojrzałym innowacją jest także zdolność zbudowania modelu bardziej audytowalnego, bezpiecznego, energooszczędnego, odpornego na manipulacje, łatwiejszego do wycofania i lepiej kalibrującego niepewność. Regulacja może w ten sposób przesunąć oś konkurencji z samej mocy modelu ku jakości governance modelu.

Od etyki dobrej woli do instytucjonalnej architektury odpowiedzialności

Najciekawszym aspektem przejścia od kodeksów etycznych do AI Act jest fakt, że prawo nie dyktuje twórcom technologii, co mają myśleć moralnie. Próbuje raczej stworzyć system, w którym moralnie istotne właściwości technologii staną się mierzalne, dokumentowalne i kontrolowalne. Ostatecznym celem nie powinno być bowiem stworzenie „etycznej maszyny" – maszyna nie jest podmiotem etycznym w takim sensie jak człowiek. Celem jest zbudowanie etycznej architektury odpowiedzialności wokół niej: obejmującej dane, model, organizację, człowieka, audyt, organ publiczny oraz prawo do zakwestionowania wyniku.

Hans Jonas pisał o konieczności rozszerzenia odpowiedzialności wraz ze wzrostem zasięgu technologicznego działania. AI Act można odczytać jako prawne urzeczywistnienie tej intuicji. Im większa zdolność systemu do wpływania na nieznane osoby, im trudniej jednostce zrozumieć jego działanie i im głębsza asymetria władzy między dostawcą a użytkownikiem, tym silniejsze stają się argumenty za odpowiedzialnością ex ante. Jednocześnie Kantowska zasada zakazująca traktowania człowieka jedynie jako środka zyskuje zaskakująco techniczny wymiar: człowiek nie może być tylko rekordem służącym optymalizacji procesu; musi pozostać podmiotem posiadającym prawa wobec tego procesu.

Rawlsowska bezstronność przeobraża się w pytanie o rozkład błędów. Proceduralizm państwa prawa ustępuje miejsca loggingowi i contestability, a etyka odpowiedzialności – risk managementowi. Filozofia zostaje przepisana na język systemu zarządzania jakością. Nie wolno jednak uznać tego za ostateczne zwycięstwo prawa nad technologią, gdyż sama regulacja może stać się rytuałem.

Można produkować tysiące stron dokumentacji, które nie zwiększają bezpieczeństwa. Można przeprowadzać audyty sprowadzone do odhaczania pól. Można stworzyć human oversight, który w rzeczywistości jest ceremonialnym klikaniem "approve". Można oznaczać deepfake, jednocześnie pozostawiając bez rozwiązania problem społecznego zaufania do autentycznych nagrań. Najbardziej niebezpieczną formą compliance jest bowiem zgodność bez refleksji. Organizacja zaczyna wtedy pytać nie o to, czy system jest dobry i bezpieczny, lecz czy posiada dokument potwierdzający spełnienie wymogów. Procedura, która miała być narzędziem odpowiedzialności, zamienia się w technikę odgradzania od niej.

Prawo potrzebuje kultury zawodowej, a kultura zawodowa prawa. Przysięga Hipokratesa dla inżynierów bez instytucji byłaby zbyt słaba, lecz instytucja pozbawiona etosu może stać się

biurokracją pozoru. Najlepszy porządek powstaje wtedy, gdy obowiązek prawny, standard techniczny i profesjonalne poczucie odpowiedzialności wzajemnie się wzmacniają.

Dochodzimy tu do bardziej uniwersalnej definicji dojrzałości technologicznej. Niedojrzała cywilizacja pyta: czy potrafimy to zbudować? Dojrzalsza dodaje: czy powinniśmy? Jeszcze dojrzalsza pyta: jakie instytucje zapewnią, że decyzja nie będzie zależała wyłącznie od dobrej woli twórcy? AI Act jest próbą odpowiedzi na to trzecie pytanie. Jednak wraz z tą regulacyjną dojrzałością pojawia się kolejny konflikt.

Jeżeli maszyna staje się coraz bardziej zdolna do wykonywania zadań poznawczych – sporządzania dokumentów, tworzenia kodu, analizy przypadków, diagnozowania, prognozowania i zarządzania informacją – odpowiedzialność za jej działanie jest tylko jedną stroną problemu. Drugą jest przemiana społecznego podziału pracy. Dotychczas technologia automatyzowała przede wszystkim siłę mięśni i powtarzalne czynności; AI zaczyna natomiast automatyzować fragmenty kompetencji, które przez stulecia stanowiły podstawę zawodowego statusu człowieka.

Czas opuścić salę regulatora i wejść na rynek pracy. Punktem wyjścia będzie brutalna metafora: dzień, w którym koń stracił pracę. Pytanie nie brzmi, czy człowiek skończy jak koń, lecz które czynności znikną, które zawody zmienią swoją wewnętrzną strukturę, kto przejmie rentę z produktywności AI i czy społeczeństwo potrafi przekształcić wzrost wydajności w nową formę ludzkiej sprawczości, zamiast jedynie w nową formę nierówności.

AI i transformacja zatrudnienia: od mechanizacji mięśni do mechanizacji kompetencji poznawczych

Jedna z najbardziej sugestywnych metafor książki Smitha prowadzi nas do Brooklynu początku lat dwudziestych XX wieku. Mechanizacja transportu i służb miejskich sprawiła, że koń stał się coraz mniej potrzebny jako jednostka napędowa gospodarki. Źródło przywołuje datę 20 grudnia 1922 roku i wycofywanie zaprzęgów strażackich jako symboliczny „dzień, w którym koń stracił pracę”, rozszerzając tę anegdotę na cały ekosystem: mniej koni oznaczało mniejszy popyt na owies, siano, stajnie, podkowy i uprzęże, a tym samym spadek zapotrzebowania na ludzi obsługujących tę infrastrukturę. Dane historyczne potwierdzają skalę tej transformacji: liczba koni i kuców na amerykańskich farmach spadła z około 19,8 mln w 1920 roku do 13,5 mln w 1930 roku, co USDA wiąże przede wszystkim z zastępowaniem zwierząt przez pojazdy silnikowe oraz mechanizację prac polowych.

Automatyzacja zadań a nie zawodów

Źródłową tezę o kaskadzie trzeba jednak naukowo zdyscyplinować. Spadek zapotrzebowania na konie niewątpliwie zmieniał strukturę rolnictwa, transportu i lokalnych rynków, ale uczynienie tej transformacji jedną z przyczyn Wielkiego Kryzysu byłoby zbyt daleko idącą inferencją bez znacznie mocniejszego aparatu historyczno-ekonomicznego. Metafora zachowuje jednak ogromną wartość, jeśli potraktować ją nie jako teorię Wielkiego Kryzysu, lecz jako model technologicznej kaskady.

Technologia rzadko usuwa tylko jedno stanowisko pracy. Zmienia ceny względne, łańcuchy dostaw, zapotrzebowanie na kompetencje, lokalizację kapitału, strukturę przedsiębiorstw i ekonomiczny sens zawodów znajdujących się kilka ogniw od bezpośrednio automatyzowanej czynności.

Pytanie o AI i zatrudnienie jest źle stawiane, kiedy brzmi: „ile zawodów zniknie?".

Zawód nie jest atomem rynku pracy. Jest wiązką zadań. Sekretarka nie jest jedną czynnością, podobnie jak lekarz, prawnik, księgowy, programista czy nauczyciel. Każdy zawód składa się z obserwacji, wyszukiwania informacji, klasyfikowania, pisania, komunikowania, negocjowania, podejmowania decyzji, odpowiedzialności, pracy fizycznej i interakcji społecznej. Technologia może przejąć część tych elementów, zwiększyć produktywność innych, uczynić niektóre niepotrzebnymi, a równocześnie stworzyć nowe.

Teoria zadań jest znacznie użyteczniejsza od katalogów „zawodów zagrożonych". Autor, Levy i Murnane pokazali już na początku XXI wieku, że komputery szczególnie dobrze zastępowały rutynowe czynności poznawcze i manualne o jasno określonych regułach, a jednocześnie uzupełniały człowieka w nierutynowym rozwiązywaniu problemów oraz zadaniach interaktywnych. Generatywna AI przesuwa jednak tę granicę. Jej szczególność polega na tym, że część czynności wcześniej klasyfikowanych jako nierutynowe poznawczo – redagowanie tekstu, tworzenie kodu, analiza dokumentu, generowanie wariantów rozwiązania czy syntetyzowanie wiedzy – zaczyna podlegać częściowej automatyzacji. Nie znaczy to jednak, że „nierutynowe" stało się nagle „automatyzowalne w całości".

Bardziej adekwatny jest obraz fragmentacji kompetencji. Model może przygotować pierwszą wersję umowy, ale prawnik odpowiada za jej sens w konkretnym stosunku gospodarczym. Może wygenerować fragment kodu, lecz ktoś musi rozumieć architekturę systemu, bezpieczeństwo i konsekwencje integracji. Może streścić dokumentację medyczną, jednak decyzja kliniczna zależy także od badania pacjenta, kontekstu, niepewności i odpowiedzialności zawodowej.

AI jako mechanizm redystrybucji kompetencji i własności wiedzy

AI nie tyle "zabiera zawód", ile przecina go wzdłuż linii zadań. Acemoglu i Restrepo proponują w tym kontekście użyteczne rozróżnienie między efektem wypierania a efektem reinstalowania pracy w nowych zadaniach. Automatyzacja przesuwa część obowiązków z człowieka na kapitał, co może obniżać udział pracy w tworzonej wartości. Przeciwwagą jest powstawanie nowych zadań, w których człowiek ponownie uzyskuje przewagę komparatywną. Ostateczny wpływ technologii na zatrudnienie zależy więc nie tylko od tempa, w jakim maszyna przejmuje nasze dotychczasowe funkcje, lecz od tego, jak szybko gospodarka generuje nowe zadania dla ludzi.

Taki podział rozbija popularną opozycję optymistów i katastrofistów. Optymista twierdzi, że każda rewolucja technologiczna ostatecznie tworzyła nowe miejsca pracy. Katastrofista odpowiada, że tym razem maszyna uczy się samego poznania, więc człowiek nie będzie miał dokąd uciec. Obie te perspektywy są zbyt ogólne. Historia nie gwarantuje automatycznie przyszłości, ale sama zdolność modelu do wykonywania wielu zadań nie dowodzi, że wszystkie funkcje gospodarcze stanowią skończony zbiór czekający na przejęcie. Gospodarka nie jest magazynem stałej liczby czynności. Gdy spada koszt jednej operacji, mogą pojawić się usługi, których wcześniej nie opłacało się świadczyć. Kiedy rachunkowość tanieje, przedsiębiorstwo może analizować więcej wariantów; gdy tłumaczenie staje się tanie, rośnie liczba komunikatów przekazywanych między językami.

Podobnie, gdy przyspiesza programowanie określonych komponentów, firma może tworzyć więcej prototypów. Technologia może zatem zmniejszać nakład pracy potrzebny do wytworzenia jednostkowego produktu, a jednocześnie zwiększać popyt na sam produkt. To klasyczny problem efektu produktywności i efektu skali. Jeśli AI obniża koszt usługi, jej cena może spaść, co z kolei zwiększy popyt. W niektórych sektorach ten wzrost może skompensować spadek zapotrzebowania na pracę przy pojedynczej usłudze; w innych nie. Nie istnieje tu uniwersalna reguła.

Zależy to od elastyczności popytu, zdolności do tworzenia nowych produktów, struktury konkurencji oraz tego, kto przejmuje nadwyżkę produktywności. Pytanie brzmi zatem nie tylko: czy AI zwiększa produktywność?, lecz: czyją produktywność, w jakim zadaniu, w jakim środowisku i komu przypada ekonomiczna renta z tego wzrostu? Empiryczne dane pokazują już dziś ogromną heterogeniczność.

W badaniu obejmującym 5172 pracowników obsługi klienta narzędzie generatywnej AI zwiększyło liczbę rozwiązanych spraw na godzinę średnio o około 15 procent. Największe korzyści odnotowali pracownicy mniej doświadczeni i słabsi, co sugeruje, że AI w pewnych obszarach działa jak

mechanizm przyspieszonego transferu wiedzy – od najlepszych praktyk do osób początkujących. Ma to ogromne konsekwencje dla ekonomii kapitału ludzkiego. Dotychczas doświadczenie było zasobem gromadzonym powoli; młody pracownik potrzebował miesięcy lub lat, by poprzez tysiące przypadków nauczyć się rozpoznawać wzorce. Jeśli model dostarcza mu wiedzy zakumulowanej przez organizację, okres dojścia do wysokiej produktywności ulega skróceniu. AI staje się wówczas protezę doświadczenia.

Może to prowadzić do kompresji różnic w wydajności między nowicjuszami a ekspertami. Z jednej strony mamy demokratyzację kompetencji; z drugiej – osłabienie premii płacowej wynikającej z posiadania specyficznej wiedzy proceduralnej. Jeżeli organizacja potrafi zakodować w modelu znaczną część metod działania najlepszych pracowników, kapitał ludzki zostaje częściowo przeniesiony z osoby do infrastruktury przedsiębiorstwa. To fundamentalna zmiana własnościowa.

Wiedza tacit knowledge, o której pisał Michael Polanyi, długo stanowiła szczególną formę przewagi pracownika: "wiemy więcej, niż potrafimy powiedzieć". Organizacja potrzebowała człowieka, ponieważ część kompetencji tkwiła w jego doświadczeniu, intuicji i pamięci przypadków. Jeśli AI potrafi wydobywać wzorce z zapisów pracy i udostępniać je innym, część tej wiedzy ulega kapitalizacji w systemie technologicznym. Wówczas może przesuwać się władza negocjacyjna.

Pracownik mniej zależy od własnej pamięci proceduralnej, ale firma mniej zależy od konkretnego pracownika. Wiedza, która kiedyś odchodziła wraz z człowiekiem po rozwiązaniu umowy, pozostaje częściowo w modelu, bazie danych, systemie workflow i dokumentacji. Cyfrowa transformacja rynku pracy jest zatem również transformacją własności wiedzy organizacyjnej. Nie we wszystkich zawodach AI spłaszcza jednak różnice kompetencyjne. W niektórych zadaniach może działać odwrotnie: ekspert lepiej wykorzystuje narzędzie, gdyż potrafi precyzyjnie sformułować problem, rozpoznać błąd i odrzucić pozornie atrakcyjną odpowiedź. Wtedy AI staje się mnożnikiem kompetencji już posiadanej. Osoba z głęboką wiedzą domenową wykorzysta model do szybszego eksplorowania przestrzeni rozwiązań, podczas gdy nowicjusz nie rozpozna fałszywego rezultatu. Powstają więc równolegle dwa mechanizmy: skill levelling, gdy AI pomaga słabszym szybciej zbliżyć się do najlepszych, oraz skill amplification, gdy najlepsi wykorzystują AI znacznie bardziej produktywnie niż reszta.

Produktywność AI to cecha systemu pracy, a nie narzędzia

To, który mechanizm dominuje, zależy od charakteru zadania. W przypadku zadań o łatwo obserwowalnym wyniku model może skutecznie prowadzić początkującego. Jednak tam, gdzie wymagana jest ocena jakości samego rozwiązania, brak wiedzy eksperckiej sprawia, że użytkownik nie potrafi zweryfikować błędu. Doskonale obrazuje to współczesne programowanie.

Wyniki badań nie układają się w prostą historię nieustannego przyspieszania pracy. W eksperymencie METR z początku 2025 roku doświadczeni programiści, pracujący nad dobrze znanymi im dużymi projektami open source, potrzebowali z użyciem ówczesnych narzędzi AI średnio o 19 procent więcej czasu. Co istotne, przed badaniem oczekiwali znacznego przyspieszenia, a po jego zakończeniu nadal subiektywnie sądzili, że pracowali szybciej.

Jest to wynik kluczowy dla psychologii organizacji. Pokazuje on, że subiektywne poczucie produktywności nie musi odpowiadać produktywności mierzonej. AI może zmniejszać wysiłek poznawczy i czynić pracę przyjemniejszą, szybko generując widoczne rezultaty, a jednocześnie tworzyć ukryte koszty weryfikacji, poprawiania i integracji. Człowiek dostrzega dziesięć sekund generowania kodu, ale znacznie gorzej zauważa kolejne minuty potrzebne do sprawdzenia, czy ten fragment pasuje do całej architektury.

Nie należy z tego badania wyciągać wniosku, że AI spowalnia programistów jako takich – sam METR podkreślał ograniczenia eksperymentu. Co więcej, w lutym 2026 roku organizacja poinformowała, że ponowienie pomiaru staje się metodologicznie trudne, gdyż część programistów nie chce już uczestniczyć w grupie kontrolnej, której zakazano korzystania z AI. Surowe dane z późniejszego okresu sugerowały możliwe przyspieszenie rzędu kilku-kilkunastu procent, jednak ze względu na silną selekcję autorzy odmówili uznania ich za solidne oszacowanie.

Właśnie to jest naukowo najciekawsze: wpływ AI na pracę jest dynamicznym parametrem, a nie stałą technologiczną. Badanie z pierwszego kwartału jednego roku może być częściowo nieaktualne rok później, ponieważ ewoluują modele, interfejsy, agenci, kompetencje pracowników i sama organizacja pracy. Technologia oraz użytkownik uczą się jednocześnie. Ekonomika AI jest zatem ekonomiką ruchomego celu.

Należy tu rozróżnić zdolność modelu do wykonania zadania od produktywności organizacji go używającej. Benchmark może wykazać, że AI potrafi rozwiązać problem programistyczny, ale przedsiębiorstwo musi ten rezultat jeszcze włączyć do kodu, przetestować, skoordynować z innymi

zespołami, zapewnić bezpieczeństwo, udokumentować decyzję i ponieść odpowiedzialność za wdrożenie.

Granica technicznych możliwości leży wcześniej niż granica ekonomicznie opłacalnego zastąpienia człowieka. Podobnie robot zdolny podnieść przedmiot nie staje się automatycznie rentownym pracownikiem magazynu. Produktywność ekonomiczna jest cechą systemu pracy, a nie pojedynczego narzędzia; obejmuje koszty integracji, nadzoru, błędów, bezpieczeństwa, szkoleń i reorganizacji procesów. Historia technologii pełna jest wynalazków, których wpływ na produktywność ujawnił się dopiero po przebudowie organizacji. Robert Solow zauważył niegdyś, że komputery można dostrzec wszędzie poza statystykami produktywności. Późniejsze badania wykazały, że przyczyną było opóźnienie organizacyjne: firma musiała nie tylko kupić komputer, ale przede wszystkim przebudować proces pracy.

AI może przechodzić podobną fazę. Dodanie chatbota do starego procesu nie jest jeszcze transformacją przedsiębiorstwa. Prawdziwa zmiana następuje wtedy, gdy proces zaczyna być projektowany od początku jako hybrydowy układ człowiek-AI. Ekonomia organizacji podpowiada w tym miejscu pojęcie komplementarności.

Przesunięcie wartości z generowania na krytyczną ocenę

Jeśli technologia przyspiesza jeden etap procesu, wąskim gardłem staje się inny. AI potrafi w kilka minut przygotować sto wariantów kampanii reklamowej, ale ktoś nadal musi je ocenić. Może zaproponować tysiące cząsteczek, lecz laboratorium ma ograniczoną zdolność ich testowania. Potrafi wygenerować więcej kodu, niż organizacja jest w stanie odpowiedzialnie zrecenzować.

Automatyzacja nie usuwa zatem wąskiego gardła; często jedynie przesuwa go dalej w łańcuchu. Stąd paradoks produktywności generatywnej AI: organizacja może zacząć produkować informacje szybciej, niż potrafi je poznawczo konsumować. Zamiast niedoboru pierwszych wersji otrzymujemy nadmiar treści wymagających selekcji. Wartość człowieka przesuwa się więc z etapu tworzenia pojedynczego materiału ku umiejętności rozpoznania, która z tysiąca propozycji jest godna działania.

W gospodarce niedoboru informacji premiowano zdolność wytworzenia odpowiedzi. W gospodarce syntetycznej obfitości coraz większe znaczenie będzie miała zdolność oceny tej odpowiedzi. Krytyczny osąd, walidacja źródeł, wykrywanie niespójności, rozumienie kontekstu i odpowiedzialny wybór zyskują na wartości właśnie dlatego, że samo generowanie staje się tanie. Z tego powodu

katalog kompetencji — ciągłe uczenie się, krytyczne rozwiązywanie problemów, komunikacja i współpraca — nie powinien być traktowany jako kolejna lista „miękkich umiejętności" dla działu HR. To raczej opis kompetencji odpornych na szybką dezaktualizację technologii. Wiedza szczegółowa pozostaje niezbędna, lecz jej wartość rośnie wtedy, gdy pozwala weryfikować i kontekstualizować wynik maszyny.

Generatywna AI może prowadzić do paradoksalnego powrotu humanistyki. Im więcej tekstu produkuje maszyna, tym cenniejsza staje się umiejętność odróżnienia argumentu od retoryki. Im więcej analiz powstaje automatycznie, tym ważniejsze jest rozumienie przyczynowości, błędu statystycznego i jakości źródła. Im łatwiej przygotować projekt regulacji, tym istotniejsza staje się znajomość historii instytucji, aksjologii i skutków niezamierzonych. Automatyzacja odpowiedzi zwiększa wartość dobrego pytania.

Warto powrócić do porównania człowieka i konia, aby wskazać jego zasadniczą granicę. Koń nie mógł zdobyć nowego zawodu, założyć przedsiębiorstwa produkującego samochody, zostać właścicielem akcji fabryki silników ani przegłosować polityki redystrybucji korzyści z mechanizacji. Człowiek może. Dlatego analogia mówiąca, że „AI zrobi z człowiekiem to, co samochód zrobił z koniem", pomija najważniejszą cechę gospodarki ludzkiej: pracownik jest zarazem konsumentem, obywatelem, twórcą nowych potrzeb, potencjalnym przedsiębiorcą i uczestnikiem instytucji. Nie oznacza to jednak, że adaptacja przebiega automatycznie czy bezboleśnie.

Makroekonomiczna równowaga nie gwarantuje indywidualnego bezpieczeństwa

Historia rewolucji przemysłowych uczy, że długookresowy wzrost dobrobytu może współistnieć z brutalnymi stratami konkretnych grup społecznych. Ekonomista, patrząc z perspektywy półwiecza, dostrzeże wyższą produktywność i pojawienie się nowych profesji. Człowiek tracący jednak zatrudnienie w wieku pięćdziesięciu lat doświadcza jednak teraźniejszości, nie długiego okresu makroekonomicznego. To właśnie tutaj pojęcie Age of Anxiety nabiera głębszego sensu. Lęk nie jest tu traktowany jako irracjonalna reakcja na postęp, lecz jako skutek dysproporcji w tempie zmian: technologia ewoluuje szybciej niż instytucje edukacyjne, systemy zabezpieczeń społecznych czy mechanizmy mobilności zawodowej.

Problem polityczny nie sprowadza się zatem wyłącznie do ostatecznego bilansu liczby miejsc pracy, lecz dotyczy czasu i geografii transformacji. Nowe stanowisko może powstać w innym mieście, sektorze czy kraju i wymagać kompetencji zupełnie innych niż te, które były potrzebne na

zlikwidowanym etacie. Statystycznie gospodarka może „stworzyć jedno miejsce za jedno miejsce", podczas gdy konkretny pracownik nadal przegrywa. To klasyczny problem niedopasowania strukturalnego. Makroekonomiczna równowaga nie jest indywidualną sprawiedliwością.

Z tego powodu najbardziej wartościowe współczesne analizy nie próbują przewidywać liczby zawodów „zabitych przez AI", lecz mierzą ekspozycję konkretnych zadań. Wspólne badanie ILO i polskiego NASK z 2025 roku, oparte na analizie niemal 30 tysięcy zadań zawodowych, szacuje, że około jedna czwarta pracowników świata wykonuje zawody wykazujące pewien stopień ekspozycji na generatywną AI. Jedynie około 3,3 procent globalnego zatrudnienia znalazło się w kategorii najwyższej ekspozycji. ILO podkreśla przy tym, że najbardziej prawdopodobnym następstwem nie jest całkowita likwidacja większości profesji, lecz ich transformacja.

Szczególnie istotne jest to, że największa ekspozycja nie dotyczy wyłącznie nisko wykwalifikowanej pracy fizycznej. Najbardziej narażone pozostają zadania biurowe, jednak wraz z rozwojem modeli rośnie ekspozycja wysoko zdigitalizowanych zawodów profesjonalnych i technicznych: analityków finansowych, programistów, specjalistów multimedialnych czy doradców inwestycyjnych. AI różni się od poprzednich fal automatyzacji właśnie tym, że wchodzi bezpośrednio do świata białych kołnierzyków. Ta zmiana fundamentalnie redefiniuje politykę społeczną.

Rewolucja robotyki przemysłowej była postrzegana głównie jako problem fabryk. Generatywna AI wprowadza niepewność do zawodów, które przez dekady uchodziły za bezpieczne dzięki wykształceniu. Dyplom nie traci na wartości, ale zmienia się jego funkcja: przestaje być gwarancją wykonywania konkretnego zestawu czynności przez czterdzieści lat, a staje się świadectwem zdolności do wielokrotnej rekonfiguracji kompetencji.

Teoria race between education and technology, kojarzona z pracami Tinbergena, Goldina i Katza, wymaga zatem aktualizacji. Przez znaczną część XX wieku wzrost podaży wykształconych pracowników pozwalał gospodarce odpowiadać na skill-biased technological change. AI tworzy jednak sytuację, w której technologia sama zaczyna przejmować czynności typowe dla osób wysoko wykształconych.

Wartość człowieka w relacyjnej odpowiedzialności

Wyścig nie toczy się już tylko między liczbą absolwentów a zapotrzebowaniem na kwalifikacje. Pytanie brzmi raczej: jakie kompetencje edukacja potrafi tworzyć szybciej, niż technologia je komodytyzuje? Może to prowadzić do głębokiej zmiany struktury prestiżu zawodowego. Umiejętność sporządzenia standardowego memorandum, prezentacji, kodu, raportu czy

tłumaczenia traci na rzadkości. Wartość przesuwa się ku zadaniom integrującym wiele domen, obarczonym wysoką odpowiedzialnością, wymagającym zaufania interpersonalnego, kontaktu z nieustrukturyzowaną rzeczywistością lub podejmowania decyzji w warunkach niepełnej wiedzy. Empatia w pielęgniarstwie i edukacji stanowi tu szczególną kategorię odporności na automatyzację. Przewaga człowieka nad AI nie wynika bowiem z prostego faktu, że „maszyna nie ma uczuć".

Istotniejsza jest społeczna relacja odpowiedzialności. Pacjent nie potrzebuje jedynie poprawnie wygenerowanych słów empatii; potrzebuje wiedzieć, że ktoś rzeczywiście bierze odpowiedzialność za jego sytuację. Dziecko nie potrzebuje wyłącznie doskonałego wyjaśnienia matematyki, lecz dorosłego, który obserwuje jego rozwój, stawia granice i uczestniczy w procesie socjalizacji. To rozróżnienie między symulacją kompetencji interpersonalnej a instytucjonalnym znaczeniem relacji będzie z czasem coraz ważniejsze. AI może znakomicie generować język wsparcia, ale nie oznacza to, że społeczeństwo zaakceptuje zastąpienie opiekuna, nauczyciela, sędziego czy lekarza przez system. W wielu zawodach wartość usługi obejmuje świadomość, że po drugiej stronie znajduje się odpowiedzialny człowiek. Jednocześnie nie wolno romantyzować „ludzkich zawodów". AI może przejąć część pracy administracyjnej pielęgniarki, uwalniając czas dla pacjenta.

Może pomóc nauczycielowi w tworzeniu materiałów dopasowanych do ucznia lub ograniczyć rutynę lekarza. Najbardziej pożądanym celem nie musi więc być obrona każdego zadania przed automatyzacją, lecz automatyzacja tego, co odbiera człowiekowi czas potrzebny na najbardziej ludzką część jego zawodu. To zasadniczo inna filozofia innowacji niż ta motywowana wyłącznie redukcją kosztów pracy. Acemoglu i Restrepo ostrzegają przed „wrong kind of AI" – kierunkiem rozwoju skoncentrowanym na zastępowaniu pracy, zamiast na tworzeniu technologii zwiększających ludzkie możliwości i generujących nowe zadania. Problem nie dotyczy zatem tylko tempa innowacji, ale jej kierunku.

Technologia nie wybiera tego kierunku samodzielnie. Robią to ceny, podatki, modele biznesowe, prawo pracy, dostępność kapitału, koszty danych, polityka przedsiębiorstw i struktura własności. Jeżeli zastąpienie pracownika jest podatkowo i organizacyjnie bardziej opłacalne niż jego technologiczne wzmocnienie, inwestycje będą podążać w stronę substytucji. Stwierdzenie, że „rynek wybrał automatyzację", często oznacza w rzeczywistości, iż konkretna konfiguracja instytucji uczyniła ją najbardziej rentownym rozwiązaniem. Pytanie o AI i zatrudnienie jest zatem również pytaniem o dystrybucję własności kapitału.

Ryzyko organizacji klepsydrowej i erozja klasy średniej

Jeśli produktywność rośnie dzięki modelom należącym do wąskiej grupy przedsiębiorstw, a pracownicy nie uczestniczą w korzyściach proporcjonalnie do tego wzrostu, gospodarka może stać się bogatsza przy jednoczesnym spadku udziału pracy w dochodzie. Wzrost PKB nie jest tożsamy z szeroko podzielonym dobrobytem. Możliwy jest paradoks: AI wykonuje więcej pracy, społeczeństwo produkuje więcej dóbr i usług, a jednocześnie rośnie niepewność ekonomiczna dużej części populacji. Problemem nie byłby wtedy niedobór produktywności, lecz mechanizm dystrybucji jej owoców. Historia gospodarcza uczy, że wzrost technologiczny i sprawiedliwość dystrybucyjna są odrębnymi zmiennymi. Stąd kluczowe znaczenie polityki przejścia.

Reskilling stał się słowem tak często nadużywanym, że grozi utratą sensu. Nie wystarczy powiedzieć zwalnianemu pracownikowi: "ucz się przez całe życie". Przekwalifikowanie wiąże się z kosztem czasu, utratą dochodu, trudnościami w opiece nad rodziną oraz ryzykiem, że nowo zdobyta umiejętność sama zostanie szybko zautomatyzowana. Polityka kompetencyjna musi więc opierać się na realnych danych o popycie na konkretne zadania, możliwościach uczenia się dorosłych i finansowaniu okresów przejściowych. Jeszcze ważniejszy może okazać się upskilling wewnątrz zawodu. Jeśli AI przede wszystkim transformuje miejsca pracy zamiast je unicestwiać, masowe przesuwanie ludzi do zupełnie nowych profesji może być mniej efektywne niż nauka nowej konfiguracji własnego zawodu. Księgowy nie musi stawać się programistą AI; może stać się księgowym potrafiącym nadzorować automatyczną analizę, wykrywać anomalie, oceniać ryzyko i komunikować konsekwencje biznesowe. Lekarz nie musi zostać data scientistem – potrzebuje rozumieć właściwości i ograniczenia narzędzi diagnostycznych.

Przyszłością pracy może więc nie być człowiek przeciwko AI, lecz człowiek z AI przeciwko człowiekowi bez AI - a następnie, na wyższym poziomie, człowiek potrafiący właściwie oceniać AI przeciwko człowiekowi jedynie naciskającemu przycisk generowania.

Podstawową nierównością przyszłości może stać się nie sam dostęp do narzędzia, lecz różnica w zdolności jego produktywnego używania. Dwa przedsiębiorstwa mogą otrzymać identyczny model, a osiągnąć zupełnie odmienne wyniki dlatego, że jedno przebudowało procesy, przeszkoliło ludzi, stworzyło mechanizmy weryfikacji i jasno podzieliło odpowiedzialność, podczas gdy drugie jedynie wykupiło licencję.

To prowadzi z powrotem do źródłowej idei People Business. Technologia nie istnieje poza ludzkim ekosystemem, a kapitał ludzki pozostaje warunkiem jej produktywnego wykorzystania. Jest to ważna korekta wobec języka "bezludnej firmy AI". Nawet bardzo zautomatyzowane przedsiębiorstwo potrzebuje ludzi projektujących cele, negocjujących kontrakty, ponoszących

odpowiedzialność, zarządzających konfliktami wartości i utrzymujących zaufanie. Możliwe jest, że największy ekonomiczny wpływ AI nie będzie polegał na masowym zniknięciu ludzi z przedsiębiorstw, lecz na gwałtownym zwiększeniu rozpiętości sprawczości pojedynczego człowieka. Jeden specjalista wspomagany agentami może wykonywać zadania, do których wcześniej potrzebował kilkuosobowego zespołu. Mała firma może uzyskać możliwości analityczne dawniej dostępne jedynie wielkiej korporacji. Przedsiębiorca może szybciej programować, tłumaczyć, projektować, analizować prawo i prowadzić marketing. Jest to demokratyzujący potencjał AI, lecz posiada on także drugą stronę.

Jeśli jeden człowiek może obsługiwać większy obszar działalności, przedsiębiorstwo może potrzebować mniejszej liczby stanowisk średniego szczebla. Powstaje ryzyko organizacji klepsydrowej: niewielka grupa wysoko kompetentnych osób sterujących dużą infrastrukturą AI oraz duża grupa pracowników usługowych i fizycznych, których automatyzacja pozostaje trudniejsza, przy jednoczesnym kurczeniu się tradycyjnych stanowisk pośrednich. Taki scenariusz miałby konsekwencje nie tylko ekonomiczne.

Klasa średnia jest kategorią dochodową, ale także instytucją społeczną. Stabilne zawody administracyjne, techniczne i profesjonalne tworzyły przez dekady ścieżki awansu, bezpieczeństwa mieszkaniowego, udziału politycznego i przewidywalności życia. Technologia zmieniająca strukturę tych zawodów może więc oddziaływać na demokrację poprzez zmianę struktury statusu społecznego.

Edukacja jako infrastruktura odporności cywilizacyjnej

Age of Anxiety nie jest w tym kontekście jedynie strachem przed robotem, lecz lękiem przed utratą życiowej mapy. Człowiek może zaakceptować konieczność ciągłej nauki, jednak znacznie trudniej funkcjonuje mu w świecie, w którym nie potrafi określić, czy inwestycja w pięcioletnie studia zawodowe zachowa wartość za dekadę. Niepewność technologiczna staje się niepewnością biograficzną.

Pytanie o przyszłość pracy znajduje swój finał nie w przedsiębiorstwie, lecz w szkole. Jeśli zawód staje się jedynie konfiguracją zmiennych zadań, edukacja nie może być rozumiana jako jednorazowe przygotowanie do z góry zdefiniowanej profesji; musi raczej kształtować zdolność do kolejnych rekonfiguracji. Nie chodzi tu o rezygnację z głębokiej wiedzy na rzecz powierzchownej „elastyczności”. Paradoksalnie w świecie AI rzetelna wiedza staje się jeszcze bardziej potrzebna, gdyż stanowi jedyny warunek weryfikacji maszyny. Wymaga to jednak nowego modelu relacji między wiedzą szczegółową a kompetencjami ogólnymi: człowiek powinien potrafić wejść głęboko

w konkretną domenę, a jednocześnie sprawnie przenosić metody rozumowania między różnymi obszarami. Można sformułować zasadniczą tezę tej części następująco: AI nie tyle odbiera człowiekowi pracę, ile destabilizuje historyczny związek pomiędzy zawodem, zadaniem, kompetencją i dochodem.

Nie wiemy jeszcze, jaki nowy układ ustabilizuje się w miejsce poprzedniego. Wiemy natomiast, że sama produktywność technologii nie rozstrzygnie, czy transformacja okaże się społecznie korzystna. Koń przegrał z silnikiem, ponieważ jego gospodarcza rola była niemal całkowicie zawarta w określonej funkcji użytkowej. Człowiek nie jest funkcją.

Potrafi on tworzyć nowe potrzeby, instytucje i znaczenia pracy, jednak ta zdolność nie jest automatyczna. Musi być podtrzymywana przez edukację, kulturę, mobilność społeczną oraz instytucje umożliwiające przejście między starym a nowym podziałem zadań. Największym błędem byłoby zatem zapewniać osobom zagrożonym transformacją, że „rynek ostatecznie stworzy nowe miejsca pracy". Zdanie to może być makroekonomicznie prawdziwe, lecz społecznie bezduszne. Instytucje muszą zająć się przestrzenią pomiędzy utratą starego zadania a możliwością wejścia w nowe – ta luka stanie się przedmiotem następnej części.

Jeśli AI przebudowuje zawody szybciej, niż tradycyjny system edukacji aktualizuje programy nauczania, szkoła, uniwersytet i kształcenie dorosłych stają się elementem infrastruktury odporności państwa. Edukacja jako infrastruktura bezpieczeństwa. Pytanie brzmi: jak zachować samodzielność poznawczą w świecie maszyn produkujących gotowe odpowiedzi?

Przejście od automatyzacji zadań do destabilizacji zawodu nie może skutkować banalnym wezwaniem, by ludzie „uczyli się przez całe życie". Takie sformułowanie jest prawdziwe, lecz intelektualnie zbyt ubogie wobec skali przemiany. Sztuczna inteligencja nie zmienia jedynie katalogu kwalifikacji poszukiwanych przez pracodawców; zaczyna oddziaływać na sam proces wytwarzania wiedzy, uczenia się, zapamiętywania, formułowania sądów i weryfikacji własnej kompetencji. Tym samym edukacja przestaje być wyłącznie sektorem usług publicznych przygotowującym jednostkę do zatrudnienia, a staje się częścią infrastruktury odporności cywilizacyjnej.

Analiza prowadzi do tego wniosku poprzez pojęcie People Business. Technologia, niezależnie od mocy obliczeniowej, pozostaje osadzona w ludzkim ekosystemie kompetencji, zaufania i instytucji. Szybkie tempo transformacji łączy się tu z „Age of Anxiety" – sytuacją, w której systemy społeczne adaptują się wolniej niż infrastruktura technologiczna. Jeżeli ta diagnoza jest trafna, szkoła i uniwersytet nie mogą ograniczać się do reagowania na listę nowych stanowisk. Muszą zwiększać zdolność społeczeństwa do absorpcji zmiany.

Wiedza faktograficzna jako niezbędne rusztowanie dla krytycznego rozumowania

W klasycznej ekonomii kapitału ludzkiego edukację opisuje się jako inwestycję zwiększającą produktywność jednostki. Becker i Schultz wykazali, że wiedza oraz umiejętności stanowią szczególny rodzaj kapitału, którego akumulacja podnosi zdolność do generowania dochodu. Perspektywa ta pozostaje użyteczna, jednak w epoce AI wymaga rozszerzenia. Skoro część wiedzy proceduralnej może zostać błyskawicznie dostarczona przez model, wartość edukacji nie może być mierzona jedynie liczbą informacji, które człowiek potrafi zreprodukować samodzielnie. Kapitałem staje się teraz umiejętność rozpoznania, czy informacja wygenerowana przez system jest prawdziwa, relewantna, wystarczająca i moralnie dopuszczalna jako podstawa działania.

Nie oznacza to jednak, że wiedza faktograficzna staje się zbędna. To jeden z najbardziej niebezpiecznych skrótów myślowych współczesnej debaty edukacyjnej: skoro maszyna potrafi w sekundę "podać odpowiedź", człowiek nie musi już niczego pamiętać. Nauki kognitywne dowodzą czegoś przeciwnego. Rozumowanie opiera się na wiedzy przechowywanej w pamięci długotrwałej; ekspert rozpoznaje istotę problemu właśnie dlatego, że posiada rozbudowane schematy pojęciowe umożliwiające interpretację nowych danych. Człowiek pozbawiony własnej wiedzy nie zyskuje wolności dzięki wyszukiwarce czy modelowi AI. Wręcz przeciwnie – staje się całkowicie zależny od jakości odpowiedzi zewnętrznego systemu, gdyż brakuje mu aparatu niezbędnego do jej weryfikacji.

Mówimy tutaj o paradoksie dostępu do wiedzy: im łatwiejsze staje się dotarcie do gotowych odpowiedzi, tym większego znaczenia nabiera wiedza pozwalająca te odpowiedzi ocenić. Człowiek nie musi pamiętać każdej daty historycznej, ale bez elementarnej chronologii nie rozpozna anachronizmu. Nie musi obliczać każdego współczynnika statystycznego ręcznie, lecz bez rozumienia prawdopodobieństwa nie zauważy, że pięknie sformatowana analiza myli korelację z przyczynowością. Nie musi znać na pamięć całego kodeksu, ale prawnik pozbawiony struktury pojęciowej prawa nie będzie w stanie rozpoznać przekonująco brzmiącej fikcji normatywnej.

Badania nad retrieval practice są w tym kontekście niezwykle niewygodne dla modnej tezy o końcu zapamiętywania. Aktywne wydobywanie wiedzy z pamięci jest jednym z najlepiej udokumentowanych sposobów wzmacniania uczenia się długotrwałego; przeglądy literatury wskazują, że efekt ten wspiera późniejsze wykorzystanie wiedzy w nowych kontekstach. W epoce AI nie powinniśmy więc usuwać pamięci z edukacji, lecz usunąć utożsamienie pamięci z całością procesu kształcenia.

Różnica jest zasadnicza. Szkoła wymagająca wyłącznie reprodukcji przygotowuje ucznia do zadań, w których maszyna ma miażdżącą przewagę. Z kolei szkoła całkowicie rezygnująca z budowania pamięci pozbawia go materiału niezbędnego do samodzielnego myślenia. Rozwiązaniem nie jest wybór między "wiedzą" a "kompetencjami", lecz wypracowanie relacji między nimi. Wiedza tworzy rusztowanie, na którym opiera się rozumowanie; rozumowanie zaś pozwala tę wiedzę przekształcać, przenosić i kwestionować.

Z punktu widzenia psychologii poznawczej kluczowe jest ograniczenie pamięci roboczej – człowiek może świadomie operować jednocześnie tylko niewielką liczbą nowych elementów. Ekspert nie posiada „magicznie" większej pamięci roboczej; dysponuje za to rozbudowanymi schematami, które pozwalają mu grupować informacje w znaczące całości. Lekarz nie widzi kilkunastu odrębnych objawów jako przypadkowego katalogu, lecz dostrzega możliwy syndrom. Szachista nie analizuje każdej figury z osobna, lecz konfigurację. Prawnik rozpoznaje typ stosunku prawnego, a historyk strukturę epoki.

Metakognicja jako tarcza przed poznawczym rozleniwieniem

AI może odciążać pamięć roboczą, lecz jeśli zostanie wprowadzone zbyt wcześnie, może uniemożliwić formowanie się niezbędnych schematów poznawczych. Uczeń, który prosi model o rozwiązanie problemu przed podjęciem samodzielnej próby, otrzymuje gotowy produkt, pomijając proces, który miał zbudować jego kompetencje. To fundamentalna różnica między wykonaniem zadania a nauczeniem się, jak je wykonać.

W tym miejscu należy rozróżnić produktywność edukacyjną od zawodowej. W przedsiębiorstwie szybkie uzyskanie poprawnej odpowiedzi jest zazwyczaj zaletą; w procesie uczenia się trudność bywa funkcjonalna. Pewne rodzaje wysiłku — samodzielne przypomnienie sobie informacji, próba rozwiązania problemu, korekta błędu czy konfrontacja kilku hipotez — stanowią mechanizm, dzięki któremu powstają trwałe struktury poznawcze. Technologie edukacyjne projektowane wyłącznie pod kątem szybkości odpowiedzi mogą zatem optymalizować wynik zadania kosztem rozwoju ucznia.

To stawia generatywną AI w specyficznej sytuacji pedagogicznej. Ten sam system może stać się protezą poznawczą albo mechanizmem poznawczego rozleniwienia. Może prowadzić ucznia poprzez serię pytań sokratejskich, dopasowywać poziom trudności czy dostarczać natychmiastową informację zwrotną, a może po prostu napisać tekst, którego uczeń nie rozumie. Narzędzie

pozostaje identyczne; różni się architektura użycia. Powraca tu główny motyw artykułu: tool versus weapon nie jest właściwością przedmiotu, lecz relacji pomiędzy przedmiotem, celem i instytucją.

W edukacji granica ta przebiega przez pojęcie metakognicji, rozumianej jako świadomość własnego procesu poznawczego i zdolność do jego regulowania. Jest to wiedza o tym, co wiem, czego nie wiem, w jaki sposób rozwiązuję problem i kiedy powinienem zmienić strategię. Badania edukacyjne od dawna wiążą rozwinięte kompetencje metakognitywne z efektywniejszym uczeniem się; wielka metaanaliza dotycząca matematyki, obejmująca 147 badań i niemal 700 tysięcy uczestników, wykazała istotny dodatni związek między metakognicją a osiągnięciami. W epoce AI metakognicja zyskuje nowy wymiar: człowiek musi kontrolować nie tylko własne rozumowanie, ale i moment delegowania tego procesu maszynie. Musi umieć odpowiedzieć sobie na pytania: czy potrzebuję teraz pomocy? Czy chcę gotowej odpowiedzi, czy jedynie wskazówki? Czy potrafię zweryfikować wynik? Czy zewnętrzne narzędzie wspiera mój proces poznawczy, czy właśnie go zastępuje?

Najnowsze prace z 2026 roku próbują łączyć ciekawość i metakognicję jako fundament autonomicznego uczenia się w dobie generatywnej AI. Autorzy zauważają, że choć modele językowe skalują możliwość indywidualnego eksplorowania pytań, ich domyślny tryb — szybkie udzielanie odpowiedzi — może osłabiać proces ciekawościowego poszukiwania, jeśli użytkownik nie potrafi świadomie regulować relacji z narzędziem. To jeden z kluczowych paradoksów przyszłej pedagogiki: najlepszy tutor AI nie zawsze powinien odpowiadać; czasami powinien umieć odmówić odebrania uczniowi problemu.

Właśnie dlatego rola nauczyciela nie znika wraz z dostępem do generatywnej wiedzy. UNESCO opisuje tę nową dynamikę jako przejście od układu nauczyciel-uczeń do relacji nauczyciel-AI-uczeń. Ramy kompetencji UNESCO dla nauczycieli obejmują pięć dziedzin: human-centred mindset, etykę AI, podstawy i zastosowania AI, pedagogikę AI oraz wykorzystanie AI w rozwoju zawodowym, definiując łącznie piętnaście kompetencji rozwijanych na poziomach od nabycia po twórcze zastosowanie.

Znaczenie tego modelu polega na odrzuceniu dwóch skrajności. Pierwsza zakłada, że nauczyciel powinien bronić szkoły przed AI; druga twierdzi, że skoro AI potrafi tłumaczyć indywidualnie, nauczyciel staje się zbędny. Obie te tezy są nietrafne. Nauczyciel przyszłości nie musi konkurować z maszyną w generowaniu encyklopedycznych odpowiedzi, lecz zarządzać procesem kształtowania autonomicznego ucznia. Wymaga to kompetencji, których nie można utożsamić z samą obsługą narzędzia. Nauczyciel musi rozumieć, kiedy użycie AI pogłębia uczenie się, a kiedy je skraca; kiedy personalizacja pomaga, a kiedy zamyka ucznia w bańce łatwych zadań; oraz kiedy automatyczna informacja zwrotna jest użyteczna, a kiedy niezbędny jest ludzki osąd dotyczący motywacji, rozwoju społecznego czy znaczenia błędu.

Smith intuicyjnie uchwycił ten problem, przywołując wcześniejsze doświadczenia z technologizacją szkół. Samo dostarczenie komputerów bez przygotowania kadry może pozostawić "martwy aktyw" — sprzęt istnieje, lecz nie tworzy wartości edukacyjnej. Źródło przeciwstawia tej logice programy takie jak TEALS i Code.org, które koncentrowały się na transferze kompetencji i budowaniu zdolności nauczycieli oraz uczniów, a nie tylko na wyposażaniu klas w urządzenia.

Ta lekcja jest w epoce AI jeszcze bardziej aktualna. Licencja na model nie stanowi polityki edukacyjnej. Posiadając najlepszy system generatywny świata, można pogorszyć proces kształcenia, jeśli zostanie on użyty do automatyzacji wszystkiego, co uczeń powinien opanować samodzielnie. Ten sam model może jednak umożliwić rzeczy dotąd kosztowne: indywidualną praktykę językową, generowanie dodatkowych problemów, symulowanie rozmów, natychmiastową korektę błędów czy wsparcie nauczyciela w pracy administracyjnej. AI nie jest zatem substytutem dydaktyki, lecz wzmacniaczem jej jakości — a co za tym idzie, może skalować zarówno dobrą, jak i złą pedagogikę.

AI literacy jako budowanie infrastruktury bezpieczeństwa epistemicznego

Podobną intuicję odnajdujemy w najnowszych ramach OECD i Komisji Europejskiej. Opublikowany w czerwcu 2026 roku AI Literacy Framework for Primary and Secondary Education definiuje AI literacy jako zestaw wiedzy, umiejętności i postaw pozwalających rozumieć działanie AI, krytycznie oceniać jej wyniki oraz korzystać z niej odpowiedzialnie i twórczo. Jeszcze rok wcześniej OECD postawiło pytanie o głębszą naturę problemu: skoro możliwości AI gwałtownie rosną, system edukacji nie może ograniczać się do dodania kilku lekcji z obsługi narzędzi, lecz musi na nowo zdefiniować kompetencje podstawowe, ich względne znaczenie oraz strukturę doświadczeń edukacyjnych. UNESCO zmierza w podobnym kierunku.

Ramy dla uczniów obejmują dwanaście kompetencji ujętych w czterech wymiarach: human-centred mindset, etyki AI, technik i zastosowań AI oraz projektowania systemów AI. Ich progresja prowadzi od rozumienia, przez stosowanie, aż po tworzenie. To istotna zmiana języka. AI literacy nie oznacza już jedynie „umiejętności wpisywania promptów” – promptowanie może być bowiem krótkotrwałą biegłością interfejsową. Prawdziwa kompetencja polega na rozumieniu systemu, jego ograniczeń, konsekwencji i kontekstu społecznego. Można to porównać do historii motoryzacji: umiejętność prowadzenia samochodu nie sprowadza się wyłącznie do naciskania pedałów, lecz obejmuje rozumienie sytuacji drogowej, ryzyka, reguł i odpowiedzialności.

Podobnie osoba kompetentna w obszarze AI nie jest tą, która najzręczniej uzyska efektowną odpowiedź, lecz tą, która wie, kiedy danej odpowiedzi nie należy użyć. Wymaga to nowej epistemologii szkolnej. Uczeń powinien być przygotowany do rozpoznawania niepewności, oceny jakości źródeł oraz rozróżniania korelacji od przyczynowości, informacji od argumentu, modelu od rzeczywistości i prawdopodobieństwa od pewności. Musi mieć świadomość, że model językowy może odpowiedzieć poprawnie bez posiadania „wiedzy" w ludzkim sensie, a jednocześnie wygenerować błąd o stylistyce eksperckiej.

Oznacza to również renesans metodologii nauki. W świecie, w którym można wyprodukować nieograniczoną ilość przekonującego tekstu, podstawowym dobrem poznawczym przestaje być informacja, a staje się nią metoda jej weryfikacji. Eksperyment, falsyfikacja, replikacja, krytyka źródła, argument logiczny, analiza statystyczna i triangulacja dowodów przestają być domeną badacza, a stają się elementami obywatelskiej higieny epistemicznej. Szkoła przyszłości powinna uczyć nie tylko tego, „jakie są fakty", lecz przede wszystkim „skąd wiemy, że są faktami". Jest to różnica między wiedzą pierwszego i drugiego rzędu. Pierwsza stwierdza: Ziemia ma określony wiek. Druga pyta, dzięki jakim metodom geologii, fizyki i pomiaru dochodzimy do takiego wniosku. W świecie syntetycznych odpowiedzi wiedza drugiego rzędu staje się tarczą przeciwko poznawczej zależności.

Jednocześnie nie wolno zamienić edukacji w permanentną szkołę podejrzliwości. Krytyczne myślenie nie polega na przekonaniu, że każdy ekspert kłamie, a każda instytucja jest skorumpowana – taki skrajny sceptycyzm prowadzi do epistemicznej fragmentacji, którą analizowaliśmy w kontekście demokracji. Dojrzała krytyczność to umiejętnością skalibrowanego zaufania: rozpoznawanie momentów, w których istnieją solidne podstawy, by zaufać wiedzy specjalistycznej, oraz tych, w których należy ją poddać dodatkowej kontroli. Podobna kalibracja powinna dotyczyć relacji z AI.

Edukacja jako infrastruktura bezpieczeństwa epistemicznego

Uczeń powinien zrozumieć, że narzędzie może być wybitne w jednym zadaniu i całkowicie zawodzić w innym; że pewność językowa nie jest pewnością epistemiczną; że wyjaśnienie stworzone post factum może brzmieć sensownie, lecz nie stanowić rzeczywistego uzasadnienia; oraz że kontekst i źródła pozostają kluczowe. Prowadzi to bezpośrednio do kryzysu tradycyjnego oceniania. Jeśli zadanie domowe sprowadza się do napisania standardowego eseju, rozwiązania typowego

problemu czy przygotowania prezentacji, generatywna AI może wykonać znaczną część produktu końcowego.

Zakaz korzystania z technologii bywa w pewnych sytuacjach pedagogicznie uzasadniony – podobnie jak uczeń matematyki powinien czasem liczyć bez kalkulatora. Nie może on jednak stanowić jedynej odpowiedzi systemowej. Ocenianie musi częściej mierzyć proces rozumowania, a nie tylko produkt. Znaczenie zyskują zatem ustna obrona rozwiązania, umiejętność wyjaśnienia metody, rekonstrukcja kolejnych wersji pracy, krytyka własnego wyniku, porównanie konkurencyjnych odpowiedzi, praca nad rzeczywistym przypadkiem oraz zdolność zastosowania wiedzy w nowej sytuacji. Nie chodzi o powrót do XIX-wiecznego egzaminu ustnego jako jedynego narzędzia, lecz o odzyskanie związku między oceną a realną kompetencją człowieka. Jeśli uczeń oddaje znakomity tekst, nie potrafiąc wyjaśnić zawartego w nim argumentu, szkoła otrzymuje dokument, a nie dowód uczenia się. To rozróżnienie istniało zawsze – plagiat również pozwalał przedłożyć cudzy tekst – jednak AI dramatycznie obniża koszt oddzielenia produktu od kompetencji.

System edukacji może zareagować na to w sposób autorytarny: monitoringiem, ciągłą kontrolą ekranów, detektorami i kulturą podejrzliwości. Byłaby to wyjątkowo ironiczna odpowiedź na technologię, która miała zwiększać wolność poznawczą. Rozsądniejszy kierunek zakłada projektowanie zadań tak, aby użycie AI nie usuwało potrzeby myślenia. Można wymagać od studenta krytyki odpowiedzi modelu, znalezienia błędów, obrony wyboru źródeł, porównania dwóch metod, wskazania obszarów niepewności lub określenia, w którym miejscu z modelu nie należało korzystać.

AI staje się wtedy przedmiotem pracy intelektualnej, a nie ukrytym ghostwriterem. Ma to dodatkową zaletę: przygotowuje człowieka do rzeczywistego środowiska zawodowego. Przyszły prawnik, lekarz, analityk czy inżynier prawdopodobnie nie będzie oceniany za to, czy potrafi ignorować AI, lecz za to, czy potrafi z niej korzystać bez utraty zawodowej odpowiedzialności.

Edukacja staje się zatem elementem wykonania zasad AI Act. Human oversight nie powstaje dzięki samemu przepisowi; pojawia się dopiero wtedy, gdy istnieje człowiek kompetentny do nadzorowania. Można prawnie nakazać, aby człowiek pozostawał w pętli, ale jeśli system edukacji nie nauczył go rozumienia niepewności, błędów modelu i odpowiedzialności za dane, nadzór będzie ceremonialny. Edukacja jest warunkiem realności regulacji technologicznej. Prawo może przyznać obywatelowi prawo do informacji, lecz ktoś musi umieć tę informację interpretować. Może zapewnić możliwość zakwestionowania decyzji algorytmicznej, ale człowiek potrzebuje aparatu intelektualnego, by rozpoznać podstawę tego kwestionowania. Może wymagać transparentności,

jednak transparentność wobec społeczeństwa pozbawionego kompetencji jest oknem wychodzącym na ścianę.

To uzasadnia określenie edukacji mianem infrastruktury bezpieczeństwa. Państwo inwestuje w cyberbezpieczeństwo, ponieważ obywatel nie jest w stanie samodzielnie bronić sieci energetycznej. Powinno zatem inwestować również w kompetencje poznawcze, gdyż demokratyczne społeczeństwo złożone z ludzi niezdolnych do odróżnienia argumentu od jego generowanej imitacji staje się podatne na manipulację – nawet wtedy, gdy jego serwery są perfekcyjnie zabezpieczone. Bezpieczeństwo państwa ma bowiem wymiar materialny, informacyjny i epistemiczny. Pierwszy dotyczy fizycznej infrastruktury, drugi integralności danych, trzeci zaś zdolności społeczeństwa do konstruowania wiarygodnej wiedzy o rzeczywistości. Szkoła znajduje się w samym centrum tego trzeciego wymiaru. Tu pojawia się problem nierówności.

Edukacja jako infrastruktura bezpieczeństwa poznawczego i nowy ideał kompetencji

Dostęp do generatywnej AI może teoretycznie demokratyzować indywidualne wsparcie edukacyjne. Uczeń z małej miejscowości zyskuje całodobowy dostęp do wyjaśnień pojęć, ćwiczeń językowych czy informacji zwrotnej. Technologia ta może częściowo obniżyć koszt tutoringu, który historycznie pozostawał przywilejem rodzin dysponujących kapitałem finansowym i kulturowym. Jednak każda demokratyzacja technologiczna niesie ryzyko powstania nowych nierówności. Pierwsza przepaść cyfrowa dotyczyła dostępu do sprzętu i Internetu; druga – jakości tego użytkowania.

W epoce AI może pojawić się trzecia przepaść: różnica w zdolności do poznawczego zarządzania maszyną. Dziecko z rodziny o wysokim kapitale kulturowym może wykorzystywać AI do dyskusji, analizy źródeł i pogłębiania wiedzy, podczas gdy uczeń pozbawiony takiego wsparcia może użyć tego samego systemu jedynie do jak najszybszego ominięcia zadania. Choć formalnie oboje mają dostęp do tej samej technologii, w rzeczywistości może ona pogłębić istniejące różnice.

Publiczny system edukacji ma zatem szczególne zadanie egalitarne. Nie może on ograniczać się jedynie do dostarczenia każdemu „AI account". Musi nauczyć każdego świadomego korzystania z narzędzia – podobnie jak szkoła nie rozwiązała problemu analfabetyzmu przez samo rozdawanie książek, lecz poprzez wieloletnią naukę czytania. Przykłady programów TEALS, Code.org oraz inicjatyw w regionach dotkniętych deindustrializacją potwierdzają tę logikę: infrastruktura technologiczna zyskuje wartość społeczną dopiero w połączeniu z transferem kompetencji i zdolnością lokalnych instytucji do ich wykorzystania.

W tym miejscu edukacja spotyka się z polityką regionalną. Transformacja technologiczna nie zachodzi bowiem równomiernie. Metropolie dysponują uniwersytetami, firmami technologicznymi i gęstym rynkiem pracy, co pozwala na szybką adaptację nowych umiejętności. Region peryferyjny może mieć dostęp do tej samej usługi AI, lecz brak mu ekosystemu pozwalającego przekuć ją w nowe zawody czy przedsiębiorstwa.

Polityka edukacyjna nie powinna zatem być domeną wyłącznie ministerstwa oświaty; jest ona elementem strategii rozwoju regionalnego, innowacyjności i odporności społecznej. Uniwersytet, szkoła zawodowa, przedsiębiorstwo i samorząd tworzą lokalny system uczenia się gospodarki. Wiedza przepływa nie tylko w podręcznikach, ale między organizacjami: pracownik zmieniający firmę przenosi praktykę, nauczyciel współpracujący z biznesem aktualizuje wiedzę, a laboratorium uniwersyteckie tworzy spin-off.

Samorząd może finansować programy odpowiadające na lokalne luki kompetencyjne. Wykorzystanie danych o brakach w zasobach talentów do kształtowania polityki edukacyjnej jest kluczowe: dzięki temu szkoła przestaje być instytucją patrzącą wyłącznie wstecz, a zaczyna reagować na realne potrzeby regionu. Należy jednak uważać, by edukacja nie została zredukowana do roli działu HR gospodarki. Jeśli ma stanowić infrastrukturę bezpieczeństwa, musi rozwijać kompetencje niemające natychmiastowej ceny rynkowej: rozumienie historii, filozofii, literatury, etyki, prawa oraz instytucji demokratycznych. W świecie AI jest to ważniejsze niż kiedykolwiek. System może szybciej wygenerować kod, ale nie odpowie na pytanie, jaki system warto kodować. Może zoptymalizować procedurę administracyjną, lecz nie ustali, jaką koncepcję sprawiedliwości ma ona realizować. Wygeneruje tysiąc wariantów komunikatu politycznego, jednak to etyka i filozofia polityczna wyznaczają granicę manipulacji.

Humanistyka przestaje być luksusem dodawanym po opanowaniu techniki, a staje się częścią systemu sterowania technologią. Im potężniejsze narzędzie, tym ważniejsza wiedza o człowieku, któremu ma ono służyć. Można to ująć w paradoksalnej formule: era sztucznej inteligencji zwiększa zapotrzebowanie na nauki o człowieku. Bez historii nie rozumiemy mechanizmów władzy; bez filozofii nie odróżniamy możliwości od powinności; bez socjologii nie dostrzegamy skutków skalowania klasyfikacji; bez psychologii nie rozumiemy automation bias; bez prawa nie przekształcimy wartości w obowiązek, a bez literatury tracimy dostęp do najbogatszego laboratorium ludzkich motywacji.

Z kolei nauki ścisłe muszą przestać funkcjonować w edukacyjnym getcie technicznym. Obywatel nie musi być zawodowym statystykiem, ale w świecie modeli predykcyjnych powinien rozumieć prawdopodobieństwo i błąd pomiaru. Nie musi być informatykiem, lecz powinien wiedzieć, że dane

treningowe nie są tożsame z rzeczywistością. Nie musi tworzyć sieci neuronowych, ale musi rozumieć, że wynik modelu zależy od danych, celu i warunków zastosowania.

Potrzebujemy nowego renesansowego ideału kompetencji – jednak nie w sensie encyklopedycznego znawcy wszystkiego, lecz jako zdolności łączenia języków różnych dziedzin. Najważniejsze decyzje dotyczące AI zapadają bowiem dokładnie na granicach: między informatyką a prawem, medycyną a statystyką, biznesem a etyką, wojskowością a filozofią oraz edukacją a psychologią.

Samodzielność epistemiczna jako fundament suwerenności człowieka

Problem dotyczy także kształcenia dorosłych. Jeśli zawód będzie coraz częściej rekonfigurowaną wiązką zadań, model edukacji oparty na schemacie „najpierw dwadzieścia lat nauki, potem czterdzieści lat pracy" staje się nierealistyczny. Potrzebujemy instytucji umożliwiających wielokrotny powrót do uczenia się bez konieczności rozpoczynania życia zawodowego od zera. Lifelong learning nie może być zatem wyłącznie moralnym obowiązkiem pracownika; musi stać się prawem infrastrukturalnym: dostępem do czasu, finansowania i programów odpowiadających realnym zmianom rynku oraz systemami uznawania krótszych form zdobywania kompetencji. W przeciwnym razie hasło „ucz się przez całe życie" stanie się formą przerzucenia kosztu transformacji technologicznej na jednostkę. Państwo, które toleruje szybkie przekształcanie struktury zatrudnienia, lecz pozostawia całe ryzyko adaptacji pracownikowi, prywatyzuje koszt innowacji. Przedsiębiorstwo osiąga produktywność, a społeczeństwo ponosi koszty przejścia.

Dojrzała polityka technologiczna powinna traktować reskilling podobnie jak infrastrukturalne inwestycje umożliwiające korzystanie z nowej technologii. Edukacja dorosłych nie może się jednak ograniczać do szkolenia z aktualnego narzędzia – konkretna platforma może zniknąć szybciej, niż system szkoleniowy zdąży wydać certyfikat. Najbardziej trwałe kompetencje dotyczą uczenia się nowego systemu, rozumienia logiki domeny, weryfikacji wyniku i komunikacji pomiędzy człowiekiem a technologią.

Najważniejszym produktem edukacji jest samosterowność poznawcza. Człowiek samosterowny to nie ten, który wszystko wie, lecz ten, który potrafi rozpoznać własną niewiedzę, znaleźć wiarygodne źródło, ocenić dowód, poprosić o pomoc, sprawdzić wynik i zmienić przekonanie pod wpływem lepszych argumentów. To właśnie metakognicja, ciekawość i samoregulacja tworzą rdzeń uczenia się zdolnego przetrwać zmianę technologiczną. AI może ten rdzeń wzmocnić albo osłabić: może

stać się cierpliwym partnerem prowadzącym człowieka przez pytania, ale może też przejąć tak wiele poznawczych mikroczynności, że użytkownik zacznie tracić zdolność do samodzielnego rozpoczęcia rozumowania.

Pytanie przyszłej pedagogiki nie brzmi: czy uczniowie powinni używać AI? To pytanie jest zbyt prymitywne. Prawdziwe pytanie brzmi: w którym momencie procesu poznawczego AI powinna wejść, w jakiej roli i kiedy powinna się wycofać, aby maksymalizować rozwój człowieka, a nie jedynie szybkość wykonania zadania? Tak postawiona kwestia przenosi nas z obszaru technologii edukacyjnej do filozofii człowieka. Edukacja nie istnieje tylko po to, by zwiększać produktywność; ma również przygotować obywatela zdolnego do wolności. A wolność w świecie maszyn generujących argumenty wymaga samodzielności epistemicznej.

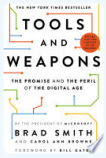
Społeczeństwo, które outsourcuje odpowiedzi, nie może outsourcować zdolności osądu. Jeśli to zrobi, AI stanie się nie tylko narzędziem pracy, ale protezą poznawczą, bez której człowiek utraci zdolność do samodzielnego poruszania się po świecie informacji. Jeśli natomiast edukacja zachowa wiedzę podstawową, metakognicję, krytykę źródeł, naukową metodę, kompetencje społeczne i zdolność formułowania problemów, technologia może odciążyć człowieka od rutyny bez odbierania mu intelektualnej suwerenności. Wtedy szkolna ławka staje się elementem bezpieczeństwa państwa równie realnym jak centrum danych. Szkoła broni społeczeństwa nie przed zagranicznym pociskiem, lecz przed czymś subtelniejszym: utratą zdolności rozumienia narzędzi, od których społeczeństwo zależy.

To prowadzi do zmiany skali. Gdy technologia kształtuje edukację, bezpieczeństwo, wymiar sprawiedliwości, komunikację i gospodarkę, przedsiębiorstwo dostarczające tę technologię przestaje być wyłącznie sprzedawcą produktu. Zaczyna negocjować z państwami, bronić klientów przed organami innych krajów, ustanawiać własne reguły etyczne, uczestniczyć w konfliktach cybernetycznych, chronić procesy wyborcze i decydować, na których rynkach może funkcjonować bez naruszania praw człowieka. Pojawia się nowy aktor polityczny, którego klasyczna teoria stosunków międzynarodowych nie przewidziała w pełni: korporacja infrastrukturalna posiadająca funkcjonalne atrybuty quasi-suwerenności.

> W ostatecznym rozrachunku największym zagrożeniem nie jest maszyna, która myśli jak człowiek, lecz człowiek, który zaczyna myśleć jak maszyna – bezkrytycznie przyjmując gotowe odpowiedzi i rezygnując z trudu samodzielnego osądu. Jeśli outsourcujemy zdolność rozumowania w zamian za wygodę szybkości, ryzykujemy stworzenie społeczeństwa, które posiada najpotężniejsze narzędzia w historii, ale utraciło mapę

> pozwalającą określić kierunek, w którym należy z nimi iść. Czy w świecie perfekcyjnie zoptymalizowanych odpowiedzi będziemy jeszcze potrafili zadać pytanie o sens?

Inspiracje

[1] 

"Tools and Weapons" Brad Smith

2019 | Penguin Press | ISBN: 978-1984877710

Bradford Lee Smith (ur. 17 stycznia 1959 r.) to amerykański prawnik i dyrektor biznesowy, który pełni funkcję wiceprezesa i prezesa firmy Microsoft. Dołączył do firmy w 1993 r., wcześniej pełnił funkcję głównego doradcy prawnego, gdzie kierował działaniami mającymi na celu rozwiązanie poważnych kontrowersji związanych z prawem antymonopolowym. Smith jest powszechnie uznawany za lidera w kluczowych kwestiach na styku technologii i społeczeństwa, w tym cyberbezpieczeństwa, sztucznej inteligencji, prywatności, praw człowieka i zrównoważonego rozwoju środowiska. Często określany jako de facto ambasador branży technologicznej, często zeznaje przed Kongresem Stanów Zjednoczonych i rządami międzynarodowymi. Smith posiada tytuł licencjata (Bachelor of Arts) Uniwersytetu Princeton oraz tytuł doktora prawa (Juris Doctor) z Columbia Law School, a także studiował prawo międzynarodowe i ekonomię w Graduate Institute w Genewie. Jest wybitnym orędownikiem społecznej odpowiedzialności biznesu w erze cyfrowej.

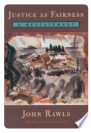
Bradford Lee Smith (born January 17, 1959) is an American attorney and business executive who serves as the Vice Chair and President of Microsoft. Joining the company in 1993, he previously served as its general counsel, where he led efforts to resolve major antitrust controversies. Smith is widely recognized for his leadership on critical issues at the intersection of technology and society, including cybersecurity, artificial intelligence, privacy, human rights, and environmental sustainability. Often described as a de facto ambassador for the technology industry, he frequently testifies before the U.S. Congress and international governments. Smith holds a Bachelor of Arts from Princeton University and a Juris Doctor from Columbia Law School, and he has also studied international law and economics at the Graduate Institute in Geneva. He is a prominent advocate for corporate responsibility in the digital age.

Microsoft

Literatura uzupełniająca

Książki

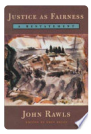
Książki autorów przywoływanych i literatura pokrewna:



[1] "Justice as Fairness"

John Rawls

Belknap Press | ISBN: 9780674005112



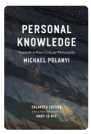
[2] "Justice as Fairness: A Restatement"

John Rawls

Harvard University Press | ISBN: 9780674005105

[3] "Humanity 'One Misunderstanding, Miscalculation Away from Nuclear Annihilation', Secretary-General Warns as Non-Proliferation Treaty Review Conference Begins"

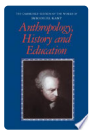
António Guterres



[4] "Personal Knowledge"

Michael Polanyi

2015 | University of Chicago Press | ISBN: 9780226232621



[5] "Anthropology, History, and Education"

Immanuel Kant

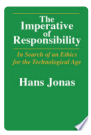
2007 | Cambridge University Press | ISBN: 9780521452502



[6] "Lectures on Ethics"

Immanuel Kant

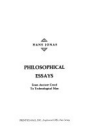
1930 | Taylor & Francis | ISBN: 9780416724806



[7] "The Imperative of Responsibility"

Hans Jonas

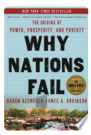
2025 | University of Chicago Press | ISBN: 9780226850337



[8] "Philosophical Essays: from Ancient Creed to Technological Man"

Hans Jonas

1974



[9] "Why Nations Fail: The Origins of Power, Prosperity, and Poverty"

Daron Acemoglu

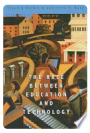
Crown Currency | ISBN: 9780307719232



[10] "Understanding the Gender Gap: An Economic History of American Women"

Claudia Goldin

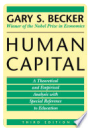
Oxford University Press, USA



[11] "The Race Between Education and Technology"

Claudia Dale Goldin, Lawrence F. Katz

2008 | Harvard University Press | ISBN: 9780674028678



[12] "Human capital: a theoretical and empirical analysis, with special reference to education"

Gary Becker

University of Chicago Press | ISBN: 9780226041223

[13] "Transforming Traditional Agriculture / Theodore W. Schultz"

Theodore Schultz



[14] "Investment in Human Capital"

Theodore William Schultz

1971 | New York : Free Press



[15] "The Art of Automation"

Jerry Cuomo, Rama Akkiraju, Allen Chan, Harley Davis, Ethan Glasman, Eileen Lowry, Rob Nicholson, Salman Sheikh

2022 | Bookbaby | ISBN: 9781667822686

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Prediction Policy Problems"

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, Ziad Obermeyer

2015 | American Economic Review | DOI: 10.1257/aer.p20151023

Źródło

[2] "Cause-Specific Risk of Hospital Admission Related to Extreme Heat in Older Adults"

Jennifer F. Bobb, Ziad Obermeyer, Yun Wang, Francesca Dominici

2014 | JAMA | DOI: 10.1001/jama.2014.15715

Źródło

[3] **"Contributions of risk factors and medical care to cardiovascular mortality trends"**
Majid Ezzati, Ziad Obermeyer, Ioanna Tzoulaki, Bongani M. Mayosi, Paul Elliott
2015 | Nature Reviews Cardiology | DOI: 10.1038/nrcardio.2015.82
Źródło

[4] **"Emergency care in 59 low- and middle-income countries: a systematic review"**
Ziad Obermeyer, Samer Abujaber, Maggie Makar, Samantha J. Stoll, Stephanie R. Kayden
2015 | Bulletin of the World Health Organization | DOI: 10.2471/blt.14.148338
Źródło

[5] **"Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability"**
Jon Kleinberg, Sendhil Mullainathan
2019 | DOI: 10.2139/ssrn.3391770
Źródło

[6] **"Language Generation in the Limit"**
Jon Kleinberg, Sendhil Mullainathan
2024 | Advances in Neural Information Processing Systems 37 | DOI: 10.52202/079017-2111
Źródło

[7] **"Author Correction: Node-based generalized friendship paradox fails"**
A. Evtushenko, J. Kleinberg
2023 | Scientific Reports | DOI: 10.1038/s41598-023-34980-5
Źródło

[8] **"Mitigating bias in algorithmic hiring"**
Manish Raghavan, Solon Barocas, Jon Kleinberg, Karen E. C. Levy
2020 | DOI: 10.1145/3351095.3372828
Źródło

[9] **"Inherent Trade-Offs in the Fair Determination of Risk Scores"**
Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan
2017 | DROPS (Schloss Dagstuhl – Leibniz Center for Informatics) | DOI: 10.4230/lipics.itcs.2017.43
Źródło

[10] "Inherent Tradeoffs in the Fair Determination of Risk Scores"

Manish Raghavan

2023 | ACM eBooks | DOI: 10.1145/3603195.3603197

Źródło

[11] "Critique of the Power of Judgment"

Immanuel Kant

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

Źródło

[12] "Critique of Judgment"

Immanuel Kant

1994 | State University of New York Press eBooks | DOI: 10.2307/jj.18254729.16

Źródło

[13] "Anthropology from a Pragmatic Point of View"

Immanuel Kant

1974 | DOI: 10.1007/978-94-010-2018-3

Źródło

[14] "Meeting the prevention challenge"

António Guterres

2017 | United Nations chronicle/UN chronicle | DOI: 10.18356/cb31790f-en

Źródło

[15] "Let us work toghether to prevent and end atrocity crimes and all affronts to human rights and human dignity"

António Guterres

2018 | Dialnet (Universidad de la Rioja)

Źródło

[16] "Struggle of the Middle East Refugees"

António Guterres

2017 | The American University in Cairo Press eBooks | DOI: 10.2307/j.ctv2ks70cq.17

Źródło

[17] "Care episode retrieval: distributional semantic models for information retrieval in the clinical domain"

Hans Jonas Fossum Moen, Filip Ginter, Erwin Marsi, Laura-Maria Peltonen, Tapio Salakoski

2015 | BMC Medical Informatics and Decision Making | DOI: 10.1186/1472-6947-15-s2-s2

Źródło

[18] "Responsibility Today: The Ethics of an Endangered Future 1"

Hans Jonas

2017 | Global Ethics and Moral Responsibility | DOI: 10.4324/9781315584829-2

Źródło

[19] "The Imperative of Responsibility: In Search of an Ethics in the Technological Age"

M. Granger Morgan, Hans Jonas

1985 | Journal of Policy Analysis and Management | DOI: 10.2307/3324683

Źródło

[20] "Models of recursion"

John Murnane

1991 | Computers & Education | DOI: 10.1016/0360-1315(91)90026-n

Źródło

[21] "The Colonial Origins of Comparative Development: An Empirical Investigation"

Daron Acemoğlu, Simon Johnson, James A. Robinson

2001 | American Economic Review | DOI: 10.1257/aer.91.5.1369

Źródło

[22] "The Colonial Origins of Comparative Development: An Empirical Investigation: Reply"

Daron Acemoğlu, Simon Johnson, James A. Robinson

2012 | American Economic Review | DOI: 10.1257/aer.102.6.3077

Źródło

[23] "Why Nations Fail: The Origins of Power, Prosperity and Poverty"

Daron Acemoğlu, James A. Robinson

2012 | Southeast Asian Economies | DOI: 10.1355/ae29-2j

Źródło

[24] **"Robots and Jobs: Evidence from US Labor Markets"**

Daron Acemoğlu, Pascual Restrepo

2019 | Journal of Political Economy | DOI: 10.1086/705716

Źródło

[25] **"Labor Productivity Functions in Meat Packing"**

Robert Solow, William H. Nicholls

1950 | The Review of Economics and Statistics | DOI: 10.2307/2937539

Źródło

[26] **"Productivity and Technical Change"**

Robert M. Solow, W. E. G. Salter

1962 | The Review of Economics and Statistics | DOI: 10.2307/1926674

Źródło

[27] **"Politics and markets—The world's political—Economic systems"**

Jan Tinbergen

1978 | Journal of Comparative Economics | DOI: 10.1016/0147-5967(78)90015-x

Źródło

[28] **"Statistical testing of business-cycle theories"**

Jan Tinbergen

1968 | Medical Entomology and Zoology

Źródło

[29] **"Macro models for developing countries"**

Jan Tinbergen

1980 | Journal of Policy Modeling | DOI: 10.1016/0161-8938(80)90008-3

Źródło

[30] **"A POSITIVE AND A NORMATIVE THEORY OF INCOME DISTRIBUTION"**

Jan Tinbergen

1970 | Review of Income and Wealth | DOI: 10.1111/j.1475-4991.1970.tb00699.x

Źródło

[31] **"The Race between Education and Technology"**

Claudia 1946- Goldin, Lawrence F. Katz

2010 | Harvard University Press eBooks | DOI: 10.2307/j.ctvjf9x5x.11

Źródło

[32] "Extending the Race between Education and Technology"

David Autor, Claudia Goldin, Lawrence Katz

2020 | DOI: 10.3386/w26705

Źródło

[33] "The Race between Education and Technology: The Evolution of U.S. Educational Wage Differentials, 1890 to 2005"

Claudia Goldin, Lawrence Katz

2007 | DOI: 10.3386/w12984

Źródło

[34] "Capital Formation by Education"

Theodore W. Schultz

1960 | Journal of Political Economy | DOI: 10.1086/258393

Źródło

[35] "Erratum to: Economics of the Family: Marriage, Children and Human Capital"

Theodore W. Schultz

1975 | Demography | DOI: 10.2307/2060726

Źródło

[36] "Knowledge, Agriculture and Welfare"

Theodore W. Schultz

1972 | Science Studies | DOI: 10.1177/030631277200200404

Źródło

[37] "The Economic Importance of Human Capital in Modernization"

Theodore W. Schultz

1993 | Education Economics | DOI: 10.1080/09645299300000003

Źródło

Artykuły naukowe autora:

[38] "Définir les règles de notre avenir numérique : l'UE est-elle sur la bonne voie ?"

Brad Smith

2021 | RED | DOI: 10.3917/red.003.0139

Źródło

[39] **"Direito e Regulação na Internet: desafios jurídicos e oportunidades para o crescimento econômico"**

Brad Smith

2012 | Law, State and Telecommunications Review | DOI: 10.26512/lstr.v4i1.21579

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: a3fa4922-69dd-4469-8228-a98439987 84e