Władza algorytmiczna i prawo do zakwestionowania maszyny: od metafory Zamku do architektury kontroli w kontekście The Hour of the Predator



AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje naturę władzy algorytmicznej, odrzucając tezę o inherentnym autorytaryzmie AI na rzecz koncepcji potencjału instytucjonalnego. Autor wskazuje, że to nie sama technologia, lecz architektura społeczno-techniczna – obejmująca dane, procedury i brak nadzoru – determinuje polityczny charakter systemów. Kluczowym elementem jest rozróżnienie czterech rodzajów nieprzejrzystości (technicznej, organizacyjnej, poznawczej i instytucjonalnej) oraz dystynkcja między wyjaśnialnością a interpretowalnością. Tekst podkreśla, że prawo do zakwestionowania decyzji maszyny nie wymaga matematycznego zrozumienia modelu, lecz przejrzystych przesłanek decyzyjnych i skutecznych ścieżek odwoławczych, co jest niezbędne dla zachowania zasad praworządności w erze automatyzacji.

The article analyzes the nature of algorithmic power, rejecting the thesis of AI's inherent authoritarianism in favor of the concept of institutional potential. The author argues that it is not the technology itself, but the socio-technical architecture—comprising data, procedures, and a lack of oversight—that determines the political character of these systems. A key element is the distinction between four types of opacity (technical, organizational, cognitive, and institutional) as well as the distinction between explainability and interpretability. The text emphasizes that the right to challenge a machine's decision does not require a mathematical understanding of the model, but rather transparent decision-making premises and effective appeal processes, which are essential for upholding the principles of the rule of law in the era of automation.

Artykuł poddaje krytyce tezę o immanentnie autorytarnym charakterze sztucznej inteligencji, argumentując, że technologia ta nie posiada własnej natury politycznej, lecz stwarza potencjał instytucjonalny do sprawowania arbitralnej władzy. Autor dowodzi, że zagrożenie nie płynie z samych algorytmów, lecz z konkretnych konfiguracji społeczno-technicznych: koncentracji mocy obliczeniowej, braku transparentności oraz iluzorycznego nadzoru człowieka (tzw. rubber-stamp oversight). Główną tezą tekstu jest konieczność przejścia od etycznych deklaracji do projektowania twardej architektury władzy, w której kluczowym bezpiecznikiem staje się 'contestability' – realne prawo do zakwestionowania decyzji maszyny. Analizując zjawiska takie jak automation bias czy odpowiedzialność rozproszona (responsibility gaps), tekst postuluje wprowadzenie zasady rozdzielenia uprawnień algorytmicznych oraz stworzenie systemowych mechanizmów kontroli, które zapobiegną sytuacji, w której prędkość automatyzacji decyzji całkowicie unieważnia ludzką zdolność do odwołania się od nich.

SŁOWA KLUCZOWE

władza algorytmiczna | prawo do zakwestionowania maszyny | contestability | automation bias | AI Act | human oversight | explainable AI (XAI) | black box | algorithmic authority | proxy discrimination | moral crumple zone | uncertainty-aware AI | probabilistyczna legitymizacja decyzji | asymetria informacyjna | separation of algorithmic powers

AI nie jest autorytarna, lecz posiada potencjał instytucjonalny

Sztuczna inteligencja, władza algorytmiczna i prawo do zakwestionowania maszyny. Od metafory "Zamku" do teorii kontroli nad decyzją

Najbardziej radykalna teza materiałów inspirowanych Giuliano da Empolim głosi, że sztuczna inteligencja jest technologią niejako z natury autorytarną: centralizuje dane i moc obliczeniową, generuje odpowiedzi, których użytkownik nie potrafi odtworzyć, a następnie domaga się przyjęcia wyniku bez pełnego wglądu w proces prowadzący do konkluzji. AI zostaje tu przyrównana do kafkowskiego "Zamku" – odległej, bezosobowej instancji, której orzeczenia wpływają na człowieka, choć ten nie potrafi wskazać odpowiedzialnego urzędnika, poznać wszystkich przesłanek ani

skutecznie odwołać się od rezultatu. Jest to metafora intelektualnie cenna, lecz jako teoria technologii zbyt kategoryczna. Sztuczna inteligencja nie posiada bowiem inherentnej konstytucji politycznej.

Ten sam rodzaj modeli może służyć scentralizowanemu aparatowi nadzoru, lekarzowi wspomagającemu diagnozę, małemu przedsiębiorstwu, badaczowi, obywatelowi czy rozproszonej wspólnocie programistów. Trafniej jest zatem twierdzić, że AI posiada autorytarny potencjał instytucjonalny, o ile zostanie połączona z koncentracją danych, brakiem kontroli, nieprzejrzystością, przymusem podporządkowania się wynikowi oraz brakiem skutecznej ścieżki zakwestionowania decyzji. Taka korekta jest istotna metodologicznie, gdyż pozwala przenieść dyskusję z poziomu moralnej natury technologii na poziom architektury władzy społeczno-technicznej. Nie pytamy już, czy algorytm "jest autorytarny", lecz kto go projektuje, kto ustala funkcję celu, na jakich danych został wytrenowany i kto wybiera próg decyzyjny. Pytamy, kto ponosi koszt błędu, kto może zakwestionować wynik oraz czy system działa jako rekomendacja, filtr, ranking czy wiążące rozstrzygnięcie. To właśnie te elementy determinują polityczne znaczenie AI.

Luciano Floridi i inni badacze filozofii informacji wielokrotnie podkreślali, że technologia nie funkcjonuje w próżni; jest elementem szerszego środowiska normatywnego, organizacyjnego i informacyjnego. Socjotechniczne podejście do AI wychodzi z podobnego założenia. NIST w swoim AI Risk Management Framework wymienia jako cechy wiarygodnych systemów między innymi rzetelność, bezpieczeństwo, odporność, rozliczalność, transparentność, wyjaśnialność, interpretowalność, ochronę prywatności oraz zarządzanie ryzykiem dyskryminacji.

Sam model nie jest więc pełnym przedmiotem oceny. Ocenie podlega system obejmujący ludzi, dane, organizację, procedury, interfejs i środowisko jego wykorzystania. W tym miejscu należy uporządkować pojęcie "czarnej skrzynki", gdyż nieprzejrzystość może oznaczać co najmniej cztery odmienne zjawiska. Pierwsze ma charakter techniczny: system jest tak złożony, że nawet ekspert nie potrafi przedstawić pełnego i intuicyjnego wyjaśnienia działania wszystkich jego parametrów. Drugie ma charakter organizacyjny: przedsiębiorstwo zna mechanizm, lecz nie ujawnia go ze względu na tajemnicę handlową, bezpieczeństwo lub przewagę konkurencyjną.

Trzecie zjawisko ma charakter poznawczy: wyjaśnienie istnieje, ale odbiorca nie dysponuje kompetencjami niezbędnymi do jego zrozumienia. Czwarte jest instytucjonalne: nikt nie zaprojektował procedury, która przekładałaby działanie modelu na uzasadnienie użyteczne w konkretnej sytuacji prawnej lub administracyjnej. Zlanie tych wszystkich rodzajów nieprzejrzystości w jedno pojęcie "black box" prowadzi do intelektualnego zamętu. Inaczej rozwiązuje się brak technicznej interpretowalności, inaczej tajemnicę handlową, a jeszcze inaczej

brak prawa do uzasadnienia decyzji. W pierwszym przypadku pomocne mogą być narzędzia XAI, testy kontrfaktyczne, analiza cech lub modele zastępcze. W drugim – obowiązki audytowe bądź udostępnienie informacji uprawnionemu organowi. W trzecim – tłumaczenie dostosowane do kompetencji odbiorcy, a w czwartym – proceduralna reforma instytucji podejmującej decyzję.

NIST rozróżnia właśnie explainability i interpretability. Wyjaśnialność odnosi się do reprezentacji mechanizmów działania systemu, interpretowalność zaś do znaczenia jego wyniku w kontekście konkretnego celu funkcjonalnego. To pozornie techniczne rozróżnienie ma głębokie znaczenie prawne. Osoba, której odmówiono świadczenia, nie potrzebuje wykładu o wszystkich warstwach sieci neuronowej.

Potrzebuje informacji, dlaczego w jej przypadku system przyczynił się do niekorzystnego rezultatu, jakie dane miały znaczenie, czy były one poprawne oraz co może zrobić, aby zakwestionować wynik. Prawo do wyjaśnienia nie powinno być zatem mylone z prawem do matematycznego zrozumienia całego modelu. Rule of law wymaga przede wszystkim, aby decyzja była dostatecznie przewidywalna, miała podstawę normatywną i mogła zostać poddana skutecznej kontroli. Najnowsza literatura dotycząca wykorzystania AI w administracji łączy nieprzejrzystość właśnie z problemem arbitralności: jeśli obywatel nie może poznać racjonalnej podstawy działania systemu i skutecznie jej zakwestionować, mechanizm prawny traci jedną ze swoich podstawowych funkcji ochronnych. Pozwala to sformułować centralną kategorię tej części: contestability, czyli możliwość zakwestionowania systemu i jego wyniku. Transparentność odpowiada na pytanie "co się dzieje?", a explainability – "dlaczego?".

Prawo do kwestionowania decyzji jako zapora przed arbitralnością AI

Contestability stawia pytanie o kluczowym znaczeniu politycznym: „co mogę zrobić, jeśli uważam, że system się myli?”. System może być przejrzysty i dobrze wyjaśnialny, a jednocześnie niedemokratyczny – w sytuacji, gdy człowiek rozumie dokładnie, dlaczego przegrał, lecz nie ma żadnej możliwości doprowadzenia do ponownej oceny. W klasycznym prawie administracyjnym odwołanie nie jest ozdobnikiem; to jedno z podstawowych urządzeń kontroli błędu. Instytucja pierwszej instancji może się pomylić, dlatego decyzję można poddać kontroli innego organu lub sądu. System algorytmiczny, generujący miliony mikrodecyzji, stawia przed prawem problem skali: pełne postępowanie odwoławcze wobec każdego wyniku może być nierealne, jednak całkowity brak możliwości kwestionowania tworzy technologiczny odpowiednik arbitralności. Konieczna staje się zatem gradacja procedur zależna od doniosłości skutku. Tu ujawnia się różnica między AI używaną

do rekomendacji filmu a AI wpływającą na rekrutację, kredyt, dostęp do edukacji, świadczenia publiczne, ubezpieczenie, wymiar sprawiedliwości czy opiekę medyczną.

Ten sam błąd statystyczny ma inną wagę społeczną. Jeśli system pomyli preferencję filmową, koszt jest trywialny. Jeżeli jednak pomyli ocenę ryzyka kredytowego albo zaklasyfikuje człowieka jako podejrzanego, konsekwencje mogą kształtować jego trajektorię życiową. Regulacja oparta na ryzyku wyrasta właśnie z tej asymetrii skutków. Akt o sztucznej inteligencji Unii Europejskiej opiera się na podobnej logice. We wrześniu 2026 roku teza, że AI pozostaje w Europie poza ramami regulacyjnymi, nie będzie już prawdziwa. AI Act jest wdrażany etapami: zakazane praktyki i część wymogów zaczęły obowiązywać wcześniej, obowiązki dotyczące modeli general-purpose AI wejdą w życie od sierpnia 2025 roku, zaś od 2 sierpnia 2026 roku zaczną działać kolejne kompetencje egzekucyjne oraz obowiązki przejrzystości. Zgodnie ze zmienionym harmonogramem reguły dla określonych systemów wysokiego ryzyka z Annex III mają być stosowane od 2 grudnia 2027 roku, a dla AI wbudowanej w regulowane produkty – od 2 sierpnia 2028 roku.

Materiał da Empoliego pozostaje więc cenny jako opis wcześniejszego poczucia bezradności regulacyjnej, ale wymaga aktualizacji o istniejącą już architekturę prawną. Co istotne, AI Act nie próbuje rozwiązać problemu za pomocą jednego uniwersalnego zakazu. Ustawodawca różnicuje zastosowania według ryzyka: zakazuje niektórych praktyk, wprowadza obowiązki dla modeli general-purpose, reguluje transparentność oraz przewiduje rozbudowane wymogi dla zastosowań wysokiego ryzyka. W nadchodzącym pełnym reżimie obowiązki te obejmą między innymi zarządzanie ryzykiem, jakość danych, logowanie, dokumentację, informację dla deployera, odpowiedni nadzór człowieka, cyberbezpieczeństwo, odporność i dokładność. Jest to odejście od fantazji, że każdy system można uczynić bezpiecznym za pomocą samego „etycznego kodeksu", na rzecz inżynieryjnego zarządzania ryzykiem. Centralną rolę zajmuje human oversight, lecz także to pojęcie wymaga demontażu. Popularna intuicja sugeruje: jeżeli decyzję algorytmu ostatecznie zatwierdza człowiek, zachowujemy ludzką kontrolę.

Fikcyjny nadzór człowieka a pułapka odpowiedzialności

Badania psychologiczne dowodzą, że rzeczywistość może wyglądać zgoła inaczej. Człowiek często staje się jedynie ceremonialnym zatwierdzającym wynik generowany przez system, co prowadzi nas do zjawiska automation bias. Jest to tendencja do nadmiernego polegania na sugestiach systemu automatycznego, szczególnie gdy użytkownik przyzwyczaił się do jego wysokiej skuteczności lub pracuje pod presją czasu. AI Act w art. 14 explicite odnosi się do konieczności uświadamiania osób sprawujących nadzór o ryzyku takiego nadmiernego zaufania. Najnowsze analizy prawno-

behawioralne podkreślają jednak, że automation bias nie jest wyłącznie błędem jednostki; wynika on z projektu interfejsu, kontekstu organizacyjnego, obciążenia pracą, reputacji systemu, motywacji oraz struktury odpowiedzialności.

Wyobraźmy sobie lekarza otrzymującego od systemu ocenę prawdopodobieństwa choroby. Formalnie zachowuje on pełne prawo do jej odrzucenia. Jeśli jednak organizacja mierzy jego wydajność, kolejka pacjentów jest długa, system w większości przypadków trafia, a sprzeciw wymaga dodatkowego uzasadnienia, „człowiek w pętli" może stać się fikcją. Człowiek nadal klika przycisk, ale kierunek decyzji w praktyce ustala maszyna. Można to określić jako rubber-stamp human oversight. System konstytucyjnie wygląda na kontrolowany przez człowieka, lecz behawioralnie człowiek pełni funkcję pieczęci. NIST już we wcześniejszych pracach dotyczących AI RMF podkreślał potrzebę badania nie tylko formalnego istnienia człowieka w pętli, lecz także tego, czy posiada on realną zdolność i zachętę do kwestionowania sugestii AI. To właśnie różnica między formalną obecnością a faktyczną sprawczością. Psychologia organizacji wskazuje dodatkowo na zjawisko moral crumple zone, opisane przez Madeleine Elish.

Gdy złożony system automatyczny zawodzi, odpowiedzialność może zostać przypisana ostatniemu człowiekowi w łańcuchu – operatorowi, który nominalnie posiadał kontrolę – mimo że jego realna możliwość przewidzenia i odwrócenia błędu była znikoma. Człowiek zostaje „strefą zgniotu" chroniącą organizację i technologię przed odpowiedzialnością. Władza AI może więc działać paradoksalnie: organizacja automatyzuje decyzję, ale zachowuje człowieka jako formalnego autora, aby utrzymać tradycyjny model odpowiedzialności. W takim układzie nie następuje prawdziwe „human-in-the-loop", lecz liability-in-the-human. Odpowiedzialność i kontrola zostają rozdzielone.

Teoria principal-agent pozwala dostrzec kolejny poziom problemu. Decydent deleguje część zadania modelowi, ponieważ ten posiada przewagę obliczeniową. Model nie jest jednak klasycznym agentem z własnymi intencjami; jest artefaktem skonstruowanym przez zespół aktorów: dostawcę danych, developera, firmę tworzącą model, deployera, integratora i administratora. Odpowiedzialność rozprasza się między wielu uczestników, przez co tradycyjne pytanie „kto zdecydował?" może nie mieć jednej intuicyjnej odpowiedzi.

Madeleine Clare Elish oraz późniejsza literatura z zakresu governance wskazują właśnie na problem responsibility gaps. Nie oznacza on, że prawnie nikt nie ponosi odpowiedzialności, lecz że potoczny model sprawstwa przestaje odpowiadać architekturze systemu. Błąd może wynikać z treningu, deploymentu, złej kalibracji, błędnej konfiguracji, distribution shift lub niepoprawnego użycia systemu w kontekście, dla którego nie był projektowany. Donald Norman w teorii projektowania zauważył, że błąd użytkownika jest często błędem systemu. Jeśli wiele osób popełnia ten sam „błąd", należy zbadać, czy interfejs nie zachęca do takiego zachowania. W przypadku AI zasada ta

jest kluczowa. Stwierdzenie, że „człowiek powinien był zauważyć, iż AI się myli”, może być zbyt łatwym przeniesieniem odpowiedzialności, jeśli system projektowano tak, by wzmacniać zaufanie i prezentować niepewny wynik z pozorem pewności.

Dlatego kalibracja jest równie ważna jak średnia trafność. System skuteczny w 95 procentach przypadków może być niebezpieczny, jeśli nie potrafi sygnalizować, w których pięciu procentach nie wie. Teoria decyzji odróżnia dokładność od kalibracji: dobrze skalibrowany model przypisujący 70-procentową pewność powinien mieć rację mniej więcej siedem razy na dziesięć. Jeśli model jest nadmiernie pewny, użytkownik może błędnie interpretować output jako wiedzę, a nie estymację. W praktyce potrzebna jest więc nie tylko explainable AI, lecz także uncertainty-aware AI. System powinien w odpowiednich zastosowaniach sygnalizować niepewność, distribution shift, niski poziom wiarygodności i przypadki wykraczające poza znane dane. W teorii uczenia maszynowego pojawiają się podejścia selective prediction i abstention: model może odmówić jednoznacznej decyzji, gdy jego niepewność jest zbyt wysoka. Paradoksalnie wartościowy system to nie ten, który zawsze odpowiada, lecz taki, który wie, kiedy powinien zamilknąć.

Ludzka psychologia komplikuje sprawę jeszcze bardziej. Obok automation bias istnieje bowiem algorithm aversion. Berkeley Dietvorst, Joseph Simmons i Cade Massey wykazali, że ludzie mogą nadmiernie tracić zaufanie do algorytmu po zaobserwowaniu błędu, nawet jeśli pozostaje on średnio trafniejszy od człowieka. Z kolei inne badania wskazują na algorithm appreciation, czyli skłonność do preferowania rady algorytmicznej w określonych warunkach. Relacja człowieka z systemem nie przebiega więc między „ślepą wiarą” a „racjonalną kontrolą”, lecz jest dynamicznym procesem kalibracji zaufania. Najlepszym celem projektowania nie powinno być zatem maksymalne zaufanie do AI, lecz appropriate reliance – poleganie na systemie wtedy, gdy ma on przewagę, i kwestionowanie go w momentach uzasadnionych wątpliwości.

Iluzja przejrzystości i pułapka autorytetu algorytmicznego

To zadanie znacznie trudniejsze niż samo zwiększenie explainability, a badania nad wyjaśnialnością dodatkowo komplikują tę prostą narrację. Eksperymenty dotyczące selekcji personelu pokazują, że przedstawienie użytkownikowi wyjaśnienia nie musi redukować nadmiernego polegania na błędnej poradzie AI. W serii prerejestrowanych badań błędna sugestia obniżała trafność ludzkich decyzji, a dodawanie różnych form wyjaśnień nie eliminowało tego efektu. Wyjaśnienie może zatem tworzyć iluzję epistemicznej kontroli: użytkownik odnosi wrażenie, że rozumie system, choć jego zdolność do identyfikowania błędnych wyników wcale się nie poprawiła.

Ujawnia to istotną różnicę między explanation as information a explanation as justification. Pierwsza forma opisuje, jak system doszedł do wyniku; druga próbuje uzasadnić, dlaczego należy go zaakceptować. Problem polega na tym, że wyjaśnienie opisowe bywa psychologicznie odbierane jako dowód poprawności. Im bardziej elegancki jest wykres lub narracja, tym większe ryzyko, że odbiorca utożsami zrozumiałość z prawdziwością. W języku epistemologii mamy tu do czynienia z pomieszaniem explanatory coherence z evidential warrant.

Dobra opowieść nie musi być prawdziwa. Ludzie od dawna wykazują tendencję do preferowania spójnych narracji, nawet przy słabym materiale dowodowym. Systemy generatywne potęgują ten problem, gdyż potrafią produkować teksty językowo płynne również wtedy, gdy zawarta w nich informacja jest błędna. Płynność staje się potencjalnym nośnikiem epistemicznego autorytetu. Daniel Kahneman opisywał zjawisko cognitive ease: treści łatwe do przetwarzania częściej wydają się nam znajome i wiarygodne. AI generująca gładką, uporządkowaną odpowiedź może zatem otrzymywać premię zaufania niezależną od rzeczywistej jakości treści. Jest to szczególnie niebezpieczne w systemach konwersacyjnych, gdzie interfejs antropomorfizuje proces – człowiek nie obserwuje tablicy prawdopodobieństw, lecz prowadzi pozorną rozmowę.

Joseph Weizenbaum dostrzegł ten mechanizm już przy ELIZA w latach sześćdziesiątych. Bardzo prosty program potrafił wywoływać u użytkowników wrażenie rozumienia. Dzisiejsze modele są jakościowo znacznie bardziej zaawansowane, lecz psychologiczny mechanizm antropomorfizacji pozostaje aktualny; wciąż przypisujemy intencje systemom reagującym w sposób społecznie czytelny. Z punktu widzenia teorii władzy prowadzi to do powstania algorithmic authority. Nie jest to autorytet oparty na urzędzie ani tradycji, lecz wynika z przekonania, że system przetworzył więcej danych i dostrzega wzorce niewidoczne dla człowieka. „Algorytm tak wyliczył” zaczyna funkcjonować jako nowa forma legitymizacji technicznej.

Max Weber rozróżniał władzę tradycyjną, charyzmatyczną i racjonalno-legalną. Algorytmiczny autorytet nie mieści się idealnie w żadnej z tych kategorii. Choć najbliżej mu do racjonalizacji, jego legitymizacja nie musi opierać się na jawnej regule prawa, lecz często na statystycznej reputacji skuteczności – system jest akceptowany, bo „zwykle trafia”. Można zatem mówić o probabilistycznej legitymizacji decyzji. W klasycznej administracji decyzja jest uzasadniana przepisem i faktami; w systemie predykcyjnym pojawia się trzeci element: podobieństwo statystyczne do wcześniejszych przypadków. Osoba może zostać oceniona nie tylko przez to, co zrobiła, ale przez to, do jakiej klasy przypadków jest probabilistycznie podobna.

To radykalnie zmienia problem sprawiedliwości. Arystotelesowska zasada równego traktowania podobnych przypadków wymaga ustalenia, co czyni je podobnymi. Model uczenia maszynowego dokonuje tej operacji poprzez reprezentację danych i funkcję straty. To, co dla człowieka jest

moralnie istotne, nie musi odpowiadać temu, co model uznaje za predykcyjnie użyteczne. Stąd problem proxy discrimination: model może nie używać jawnie cechy prawnie chronionej, a mimo to wykorzystywać inne zmienne silnie z nią skorelowane. Kod pocztowy może pełnić funkcję proxy dla struktury społeczno-ekonomicznej, historia zatrudnienia dla opieki nad dziećmi, a wzorce zakupowe dla zdrowia lub statusu ekonomicznego.

Usunięcie chronionej cechy nie gwarantuje więc neutralności. Jednocześnie fairness itself jest wielowymiarowe. Informatyka zna wiele formalnych definicji fairness – demographic parity, equalized odds czy predictive parity – które przy określonych różnicach bazowych nie mogą być spełnione równocześnie. Oznacza to, że „bezstronny algorytm" nie jest kategorią czysto techniczną. Wybór metryki wymaga decyzji normatywnej o tym, jaki rodzaj błędu i nierówności uznajemy za ważniejszy. AI nie usuwa zatem polityki; ukrywa politykę wewnątrz funkcji celu.

Władza AI to polityka błędu i kontrola nad infrastrukturą

Kto wybiera kompromis między false positives a false negatives, pośrednio decyduje o tym, kto ponosi ryzyko. Model wykrywający oszustwa może zostać ustawiony agresywnie, by zatrzymywać więcej rzeczywistych nadużyć, lecz ceną za to będzie częstsze blokowanie niewinnych użytkowników. Ten sam problem widzieliśmy w makroskali w Salwadorze: maksymalizacja czułości kontroli nieuchronnie zwiększa koszt fałszywie pozytywnych klasyfikacji. AI jest więc maszyną statystyczną zanurzoną w polityce błędu. Kluczowe pytanie nie brzmi zatem „jak dokładny jest model?", lecz komu wolno się pomylić i kto płaci za tę pomyłkę. W medycynie false negative może oznaczać niewykrytą chorobę, a false positive – niepotrzebny zabieg; w wymiarze sprawiedliwości proporcje moralne są już zupełnie inne.

W reklamie sytuacja wygląda jeszcze inaczej. Nie istnieje jeden uniwersalny próg. Twierdzenie Bayesa pomaga dostrzec strukturę problemu, ale nie rozwiązuje kwestii normatywnych. Probabilistyka potrafi obliczyć prawdopodobieństwo, lecz nie ustala wartości kosztu błędu – to społeczeństwo, prawo i instytucje muszą zdecydować, jak ważyć konsekwencje. Da Empoli wiąże AI z centralizacją także z innego powodu: rozwój najbardziej zaawansowanych modeli wymaga ogromnej infrastruktury obliczeniowej. Jego intuicja znajduje silne wsparcie empiryczne, choć wymaga aktualizacji.

AI Index 2026 wskazuje, że w 2025 roku przemysł stworzył ponad 90 procent „notable AI models", przy czym najbardziej zaawansowane z nich są jednocześnie najmniej transparentne pod względem danych treningowych, parametrów i kosztów. Globalna moc obliczeniowa AI rosła w tym czasie bardzo szybko, a infrastruktura pozostaje skoncentrowana wokół ograniczonej liczby dostawców

sprzętu i hyperscalerów. Tworzy to realną barierę wejścia do najbardziej kapitałochłonnej warstwy frontier AI. Ekonomia skali jest tutaj brutalna. Trenowanie modelu frontier wymaga nie tylko algorytmu, lecz ogromnych klastrów akceleratorów, centrów danych, energii, specjalistycznego personelu i kapitału. Im wyższy koszt stały, tym łatwiejsza oligopolizacja. Klasyczna teoria organizacji przemysłowej podpowiada, że wysokie koszty wejścia ograniczają liczbę efektywnych konkurentów nawet bez antykonkurencyjnej zmowy. Nie oznacza to jednak, że AI jako całość jest monopolistyczna, gdyż obok modeli zamkniętych rozwijają się modele open-weight.

OECD podkreśla, że określenie „open source AI" jest nieprecyzyjne, ponieważ otwartość może dotyczyć różnych komponentów: wag, kodu, danych, dokumentacji czy procesu treningu. Open-weight oznacza przede wszystkim publiczną dostępność wag modelu, a nie pełną transparentność całego systemu. Jest to rozróżnienie kluczowe, gdyż decentralizacja jednego elementu nie oznacza automatycznie decentralizacji całej infrastruktury. Modele open-weight mogą obniżać bariery dostępu do technologii, umożliwiać lokalne wdrożenia, audyt, dostrajanie i eksperymentowanie poza kontrolą jednego dostawcy. Mogą więc być instrumentem dystrybucji sprawczości technologicznej. Jednocześnie otwartość może utrudniać późniejsze cofnięcie niebezpiecznej zdolności, ponieważ model po rozpowszechnieniu przestaje być pod kontrolą producenta. OECD analizuje zatem otwartość jako kontinuum korzyści i ryzyk, a nie prostą dychotomię „otwarte dobre – zamknięte złe".

Pokazuje to, że centralizacja AI przebiega na kilku różnych poziomach. Wiedza naukowa może być rozproszona, wagi niektórych modeli otwarte, a jednocześnie produkcja najnowocześniejszych chipów, dostęp do ogromnych klastrów compute i usługi chmurowe mogą pozostawać silnie skoncentrowane. Skoro AI Index 2026 wskazuje na gwałtowny wzrost globalnej mocy obliczeniowej (średnio 3,3 raza rocznie od 2022 roku) przy jednoczesnej dominacji kilku dostawców sprzętu, prawdziwe pytanie nie brzmi „czy AI jest scentralizowana?", lecz która warstwa stosu technologicznego podlega centralizacji i jak bardzo jej kontrola przekłada się na władzę nad pozostałymi poziomami.

W tym miejscu pomocna jest teoria stosu (stack). Na samym dole znajdują się energia, półprzewodniki i centra danych; wyżej chmura, następnie modele fundamentowe, a potem API, aplikacje, agenci i interfejsy; na szczycie zaś konkretne zastosowania społeczne. Konkurencja może być intensywna na jednej warstwie i niemal nieobecna na innej. Państwo regulujące wyłącznie poziom aplikacji może zatem przeoczyć głęboką koncentrację infrastrukturalną.

Benjamin Bratton używał metafory The Stack do opisu planetarnej infrastruktury obliczeniowej, w której warstwy techniczne tworzą nowe formy organizacji geopolitycznej. Nawet bez przyjmowania całej jego teorii warto wykorzystać podstawową intuicję: władza cyfrowa jest wielowarstwowa i nie

zawsze znajduje się tam, gdzie użytkownik widzi interfejs. Da Empoli ma zatem rację, że AI może centralizować władzę, ale dzieje się to pod określonymi warunkami instytucjonalnymi. Jeśli dostęp do compute, danych i modeli jest silnie skoncentrowany, a państwa i przedsiębiorstwa stają się zależne od kilku dostawców, rośnie infrastructural power tych podmiotów. Jeśli jednak rozwijają się alternatywne modele, interoperacyjność, otwarte wagi, lokalne wdrożenia i konkurencyjne chmury – pojawiają się realne mechanizmy decentralizacji.

Ryzyka operacyjne AI wynikają z rutynizacji i deskilling-u, a nie z natury algorytmu

Technologia sama w sobie nie rozstrzyga wyniku. Równie błędne byłoby twierdzenie, że AI „działa tylko przez korelację" i z tego powodu nie potrafi wspierać rozumowania przyczynowego. Współczesne systemy uczenia maszynowego faktycznie często opierają się na wzorcach statystycznych, lecz mogą być łączone z modelami przyczynowymi, symulacjami, narzędziami symbolicznymi oraz eksperckimi procedurami weryfikacji. Judea Pearl konsekwentnie podkreślał różnicę między korelacją, interwencją a kontrfaktycznością, wykazując, że predykcja obserwacyjna nie jest tożsama z wyjaśnieniem przyczynowym. Właściwym zarzutem nie jest zatem brak znajomości przyczyn przez AI, lecz pytanie, czy konkretne zastosowanie wymagające rozumowania przyczynowego wykorzystuje metodę, która rzeczywiście takie wnioskowanie uzasadnia. W medycynie model może doskonale przewidywać zgon na podstawie danych historycznych, a jednocześnie błędnie wspierać decyzję terapeutyczną, jeśli nauczył się korelacji będących jedynie skutkiem wcześniejszych praktyk klinicznych. Predykcja odpowiada na pytanie „co zwykle następuje?", podczas gdy decyzja często wymaga odpowiedzi na pytanie „co nastąpi, jeśli zrobimy X zamiast Y?". To dwa odmienne problemy. James March pozwala uchwycić tutaj jeszcze jeden wymiar.

Organizacja wdrażająca AI zmienia równowagę między eksploracją a eksploatacją. Model może niezwykle sprawnie eksploatować wzorce obecne w danych historycznych, lecz właśnie dlatego może utrwalać przeszłość. Jeśli środowisko ulega zmianie, dane historyczne przestają być pełnym przewodnikiem – to właśnie problem distribution shift. Model działa najlepiej wtedy, gdy przyszłość statystycznie przypomina przeszłość, na której go trenowano. Kryzys gospodarczy, pandemia, wojna, przełom technologiczny czy reforma prawa mogą zmienić rozkład danych. W takim przypadku model nie musi „zepsuć się" technicznie; może nadal działać zgodnie z projektem, ale jego wcześniejsze relacje predykcyjne tracą aktualność. Karl Popper zauważyłby, że wiedza rozwija się poprzez możliwość falsyfikacji. System AI wdrażany w organizacji potrzebuje

analogicznego mechanizmu: ciągłego monitorowania, czy hipotezy statystyczne, na których opiera się model, pozostają aktualne. Governance modelu nie kończy się w dniu deploymentu. Potrzebny jest post-deployment epistemology – system badania, czy model nadal wie to, co organizacja sądzi, że on wie. NIST traktuje monitoring i zarządzanie ryzykiem w całym cyklu życia jako kluczowy element wiarygodności AI. W regulacji technologicznej oznacza to przejście od statycznej certyfikacji do ciągłego nadzoru. Tradycyjny produkt po certyfikacji pozostaje bowiem względnie niezmienny.

Model może natomiast ulec aktualizacji, dostrojeniu, zmianie danych, integracji z nowym narzędziem lub zmianie sposobu użycia. Tu pojawia się problem model drift oraz purpose drift. Pierwszy dotyczy spadku statystycznej wydajności; drugi – przesunięcia zastosowania poza pierwotny cel. System zaprojektowany jako asystent może z czasem stać się selektorem. Narzędzie doradcze może przeobrazić się w quasi-decyzyjne nie poprzez formalną zmianę dokumentacji, lecz dlatego, że pracownicy coraz rzadziej podważają jego sugestie. Instytucjonalne znaczenie AI nie wynika więc tylko z jej możliwości technicznych, ale z rutynizacji użycia. Anthony Giddens przypominał, że struktury społeczne reprodukują się poprzez codzienną praktykę. Jeśli tysiące pracowników każdego dnia konsultują system przed podjęciem decyzji, AI stopniowo staje się elementem struktury organizacyjnej, nawet jeśli formalnie pozostaje „tylko narzędziem".

Może to zmieniać również dystrybucję wiedzy zawodowej. Jeżeli młodszy pracownik coraz częściej otrzymuje gotową podpowiedź, może nabywać mniej kompetencji poprzez samodzielne rozwiązywanie problemów. Automatyzacja tworzy wtedy paradoks kompetencyjny: im bardziej system przejmuje rutynę, tym mniej okazji ma człowiek do ćwiczenia umiejętności niezbędnych właśnie wtedy, gdy system zawiedzie. Lisanne Bainbridge opisała już w 1983 roku „ironie automatyzacji": automatyzacja usuwa człowieka z normalnego toku pracy, pozostawiając mu jedynie najbardziej nietypowe i trudne sytuacje awaryjne – dokładnie te, w których potrzebuje doświadczenia, którego wcześniej nie mógł zdobyć przez samą obecność automatyzacji. AI jest kolejną odsłoną tego problemu. W lotnictwie autopilot poprawił bezpieczeństwo, ale wywołał pytania o utrzymywanie umiejętności manualnych. W medycynie system diagnostyczny może zwiększać trafność, lecz długofalowo zmieniać sposób, w jaki lekarze uczą się rozpoznawać wzorce. W administracji AI może przyspieszyć analizę dokumentów, ograniczając jednocześnie zdolność urzędników do samodzielnego rekonstruowania sprawy. Zyskujemy produktywność, ryzykując deskilling. Harry Braverman analizował deskilling w kapitalistycznej organizacji pracy jako przejmowanie wiedzy pracownika przez system zarządzania.

AI nadaje tej teorii nową formę. Wiedza może zostać wyekstrahowana z praktyki wielu ludzi, zakodowana w modelu i następnie zwrócona pracownikowi jako usługa, nad którą nie ma on

kontroli. Jednocześnie możliwy jest proces odwrotny: upskilling przez augmentację. AI może udostępnić małemu przedsiębiorcy zdolności analityczne wcześniej zarezerwowane dla dużych organizacji, pozwolić lekarzowi szybciej analizować literaturę, programiście pracować w nowych językach, a osobie z niepełnosprawnością korzystać z interfejsu wcześniej niedostępnego. Technologia może zatem zarówno centralizować, jak i demokratyzować kompetencje. Ta dwoistość pokazuje, dlaczego pojęcie „autorytarnej natury AI" jest naukowo nieadekwatne. Autorytarny lub emancypacyjny charakter może mieć układ instytucjonalny wykorzystujący AI, a nie abstrakcyjny algorytm. Najbardziej jaskrawą korektę należy przeprowadzić w odniesieniu do bezpieczeństwa modeli OpenAI z 2024 roku. Notatka twierdzi, że model z jesieni 2024 roku osiągnął „najwyższy poziom ryzyka" w zakresie broni biologicznej, chemicznej lub nuklearnej i mimo to został udostępniony. Oficjalna karta systemowa OpenAI o1 tego nie potwierdza.

Adaptacyjne zarządzanie AI musi łączyć elastyczność z ochroną praw podstawowych

Model został sklasyfikowany ogólnie jako Medium w kategorii CBRN, a nie High czy Critical; podobnie w badaniach nad obszarem radiologicznym i nuklearnym końcowe oceny mieściły się w poziomach Low lub Medium. W ówczesnym Preparedness Framework do wdrożenia dopuszczano modele z poziomem post-mitigation Medium lub niższym. Sensacyjna wersja źródła miesza więc pojedyncze wskaźniki testowe, scenariusze przed mitigacjami i końcową ocenę ryzyka. Nie oznacza to jednak, że problem CBRN jest banalny. Przeciwnie, system cards pokazują, że laboratoria same traktują go jako kategorię wymagającą systematycznej ewaluacji i zabezpieczeń. Właściwy wniosek nie brzmi zatem: "wypuszczono system sklasyfikowany jako najwyższe ryzyko", lecz raczej: frontier AI osiąga poziom, przy którym zdolności potencjalnie związane z niebezpiecznymi domenami wymagają formalnych progów deploymentu, testów i mitigacji. Jest to teza mniej widowiskowa, lecz naukowo mocniejsza. Dzisiejsza architektura europejska wyraźnie uwzględnia podobne kategorie. Od sierpnia 2026 roku egzekwowane są obowiązki dotyczące modeli GPAI, w tym wobec najbardziej zaawansowanych systemów mogących generować ryzyko systemowe; Komisja wskazuje wprost między innymi ryzyka CBRN, cyberofensywne, utraty kontroli, szkodliwej manipulacji i naruszeń praw podstawowych.

Ponownie oznacza to odejście od obrazu państwa całkowicie bezradnego. Nie znaczy to jednak, że regulator rozwiązał problem. AI rozwija się szybciej niż proces legislacyjny, a pomiar nowych zdolności staje się coraz trudniejszy. AI Index 2026 wskazuje na szybkie nasycanie benchmarków i problemy z jakością samych ewaluacji; niektóre testy, które miały mierzyć postęp przez lata, stają

się zbyt łatwe w ciągu miesięcy. Jeżeli miara przestaje rozróżniać modele, regulator i laboratorium tracą instrument orientacji. Prawo Goodharta powraca więc również w AI safety: gdy benchmark staje się celem, system może być coraz lepiej dostrajany do konkretnego testu, a wynik przestaje rzetelnie reprezentować szerszą zdolność. Bezpieczeństwo wymaga zatem portfela testów, red teamingu, obserwacji deploymentu i ciągłego aktualizowania metod ewaluacji. To kolejny argument za traktowaniem AI governance jak nauki eksperymentalnej, a nie jednorazowego aktu regulacyjnego. Nie istnieje bowiem ostateczna lista zagrożeń.

Zmieniają się modele, użytkownicy, integracje i sposoby wykorzystania technologii. Prawo powinno definiować odpowiedzialność i granice, ale instytucje kontrolne potrzebują zdolności adaptacyjnego uczenia. Donald Schön nazwałby to reflective practice. Instytucja nie tylko stosuje wiedzę, lecz obserwuje skutki własnego działania i na tej podstawie aktualizuje praktykę. W AI governance jest to szczególnie istotne, ponieważ część problemów ujawnia się dopiero po masowym wdrożeniu. Jednocześnie należy unikać przeciwnego błędu – nieskończonego eksperymentowania na obywatelach bez ponoszenia odpowiedzialności. Prawo badań z udziałem ludzi wprowadziło zasadę informed consent właśnie dlatego, że wiedza nie może być zdobywana kosztem jednostki bez wyznaczenia granic. Systemy cyfrowe często generują eksperymenty behawioralne na ogromną skalę, podczas gdy człowiek nie zawsze wie, że stał się częścią procesu optymalizacji.

Polityka AI musi zatem pogodzić adaptive governance z rights-preserving governance. Instytucje muszą się uczyć, ale prawa podstawowe nie mogą być traktowane jako zmienne testowe. Tu ponownie wchodzi teoria capabilities Amartyi Sena. AI należy oceniać nie tylko przez wydajność systemu, lecz przez to, jak zmienia realne możliwości działania człowieka.

Czy zwiększa dostęp do wiedzy? Czy umożliwia osobie słabiej wykształconej korzystanie z informacji? Czy ogranicza zależność od eksperta? Czy przeciwnie – uzależnia człowieka od systemu, którego nie rozumie i nie może opuścić? Wolność wobec AI powinna być więc rozumiana pozytywnie i negatywnie. W sensie negatywnym oznacza ona ochronę przed bezpodstawnym automatycznym rozstrzygnięciem. Pozytywnie – zdolność korzystania z technologii dla własnych celów, rozumienia jej ograniczeń i świadomego wybierania alternatywy. Isaiah Berlin ostrzegał przed przekształcaniem pozytywnej wolności w paternalistyczne przekonanie, że ktoś inny wie lepiej, czego jednostka naprawdę potrzebuje. Algorytmiczny paternalizm może przyjąć podobną formę: system przewiduje, co jest "najlepsze", a człowiek stopniowo traci możliwość wyboru opcji uznanej przez model za gorszą. W recommender systems jest to problem łagodniejszy; w administracji lub medycynie może być znacznie poważniejszy.

Prawo do przerwania automatyczności jako zapora przed arbitralną dominacją

Projektowanie AI powinno zatem zakładać możliwość meaningful override, a nie jedynie implementację formalnego przycisku. Pojęcie „manual override", które u da Empoliego służyło jako symbol drapieżnej polityki, zostaje tutaj intelektualnie odwrócone: w relacji człowiek-AI może ono stać się instrumentem ochrony autonomii. Paradoks polega na tym, że ten sam termin oznacza dwie przeciwstawne rzeczy, zależnie od kierunku relacji władzy. Przywódca wykorzystuje manual override, by ominąć ograniczenia instytucjonalne; obywatel lub profesjonalista potrzebuje go natomiast po to, by nie zostać uwięzionym w automatycznym systemie. Ta symetria jest jednym z kluczowych odkryć niniejszego artykułu. Władza nie spoczywa bowiem w posiadaniu automatyzacji, lecz w prawie do ostatecznego zatrzymania procesu – podobnie jak w lotnictwie istnieją mechanizmy przejęcia kontroli, a w bezpieczeństwie przemysłowym stosuje się emergency stop.

W systemie konstytucyjnym rolę tę pełnią sądy, veta i prawa podstawowe. W obszarze AI niezbędny jest ich funkcjonalny odpowiednik, który możemy nazwać prawem do przerwania automatyczności. Nie oznacza to konieczności ręcznego zatwierdzania miliardów poprawnych mikrooperacji, lecz wymóg, by system nie był zaprojektowany w sposób uniemożliwiający ludzką rewizję każdej istotnej decyzji, gdy stawką są prawa lub poważne interesy jednostki. Prawo do przerwania automatyczności wymaga jednak kompetentnego operatora; osoba pozbawiona wiedzy nie jest bowiem realnym kontrolerem. Wymaga ono również czasu. Jeśli organizacja narzuca pracownikowi normę pięciuset decyzji dziennie, rzeczywisty oversight nie powstanie poprzez samo dopisanie obowiązku „sprawdzenia wyniku AI". Jest to problem organizacyjny, a nie wyłącznie technologiczny. Socjotechniczna kontrola AI musi zatem obejmować projekt stanowiska pracy, wynagrodzenie za ostrożność, możliwość sprzeciwu, audyt ścieżki decyzji oraz analizę statystyki override'ów. Jeśli rekomendacje modelu nigdy nie są odrzucane, może to świadczyć o jego perfekcji, lecz prawdopodobniej powinno skłonić do pytania o rzeczywistą niezależność nadzoru.

Właśnie dlatego AI governance staje się w istocie governance ludzi używających AI. Sprawnego systemu nie da się zapewnić wyłącznie poprzez parametry techniczne modelu; niezbędne są normy organizacyjne, kultura krytyki, kompetencje i odpowiedzialność. Amy Edmondson w badaniach nad psychological safety dowodziła, że organizacje uczą się efektywniej, gdy ludzie mogą bezpiecznie zgłaszać błędy, wątpliwości i niezgodę. W środowisku AI zasada ta zyskuje na znaczeniu: pracownik musi mieć możliwość stwierdzenia, że „model się myli", nawet jeśli system cieszy się imponującą reputacją. W przeciwnym razie algorithmic authority zaczyna przekształcać

się w epistemic domination. Miranda Fricker pisała o epistemic injustice w relacjach międzyludzkich; analogicznie można analizować sytuację, w której indywidualne świadectwo człowieka zostaje systematycznie uznane za mniej wiarygodne niż rekord w bazie lub predykcja modelu. Gdy obywatel mówi „to nie ja", a system odpowiada wysokim prawdopodobieństwem, problemem przestaje być trafność predykcji, a staje się nim rozkład ciężaru dowodu. Prawo proceduralne musi zatem jasno określić, kto ma udowodnić błąd.

Jeśli całe ryzyko zostaje przeniesione na jednostkę pozbawioną dostępu do danych i modelu, możliwość odwołania staje się nominalna. Contestability wymaga symetrii proceduralnej wystarczającej do prowadzenia realnego sporu. Frank Pasquale w „The Black Box Society" argumentował, że władza algorytmiczna rośnie tam, gdzie instytucje wykorzystują niewidoczne systemy reputacyjne i klasyfikacyjne. Najskuteczniejszym lekarstwem nie jest pełna jawność wszystkiego, lecz asymetryczna transparentność: instytucja posiadająca władzę powinna podlegać większej kontroli niż jednostka, której dotyczy decyzja. Jest to klasyczna intuicja republikańska Philipa Pettita: wolność oznacza nie tylko brak aktualnej ingerencji, ale przede wszystkim brak arbitralnej dominacji. Człowiek nie jest w pełni wolny, jeśli jego los zależy od cudzej woli, nawet gdy ta chwilowo mu sprzyja. W systemie algorytmicznym analogiczny problem pojawia się, gdy jednostka zależy od systemu, którego decyzji nie może skutecznie zakwestionować. Republikańskie pojęcie non-domination może zatem okazać się trafniejsze niż prosty język prywatności. Prywatność pyta o to, jakie dane system posiada; non-domination pyta, jaką władzę nad człowiekiem daje posiadanie i użycie tych danych.

Rozdzielenie uprawnień jako zapora przed arbitralnością

To prowadzi do kwestii asymetrii informacyjnej analizowanej przez George'a Akerlofa. Użytkownik często nie wie, jak model działa, podczas gdy dostawca nie zna wszystkich skutków jego użycia. AI tworzy zatem podwójną niepewność: użytkownik jest słabszy informacyjnie względem dostawcy, lecz sam dostawca również nie posiada pełnej wiedzy o emergentnym zachowaniu systemu. To właśnie odróżnia AI od klasycznej maszyny.

Producent śrubokręta zna mniej więcej przestrzeń możliwych zastosowań swojego narzędzia. Dostawca modelu general-purpose nie kontroluje natomiast wszystkich promptów, integracji i kontekstów. Regulacja musi więc dzielić odpowiedzialność między providera a deployera. Taką logikę przyjmuje AI Act, zwłaszcza w zakresie przyszłych obowiązków dotyczących systemów wysokiego ryzyka oraz obecnych wymogów GPAI. Odpowiedzialność powinna zatem podążać za

kontrolą nad ryzykiem: provider odpowiada za to, co może kontrolować na etapie projektowania modelu; deployer – za kontekst wdrożenia, dane lokalne, procedury ludzkiego nadzoru i sposób wykorzystania. Użytkownik nie powinien być obciążany błędem, którego źródło leży w warstwie niedostępnej jego kontroli.

Ekonomia prawa Guido Calabresiego uczy, że odpowiedzialność warto lokować tam, gdzie można najtaniej zapobiegać szkodzie. Cheapest cost avoider w AI może być inny dla każdego rodzaju ryzyka. Model provider może najlepiej ograniczyć pewne podatności techniczne, deployer – błędne wykorzystanie, organizacja – presję na automatyczne zatwierdzanie, a regulator – brak minimalnego standardu. Nie istnieje zatem jeden „winny AI", lecz łańcuch przyczynowy i kontrolny, który dobry system odpowiedzialności powinien odwzorowywać.

Najważniejsza różnica między wizją da Empoliego a proponowanym tutaj modelem polega na tym, że „Zamek" nie musi być nieprzenikniony. Można projektować drzwi, procedury i zewnętrzne punkty kontroli. NIST, AI Act, standardy audytu, red teaming, dokumentacja modeli oraz prawa proceduralne są próbami zbudowania takich wejść. Nie dają one jednak gwarancji bezpieczeństwa. Instytucja regulacyjna może nie nadążać, audyt może stać się rytuałem compliance, explainability może produkować iluzję zrozumienia, a human oversight – być jedynie rubber stampem.

Open weights mogą decentralizować władzę, ale jednocześnie decentralizują ryzyko. Każde zabezpieczenie ma swoje failure modes. Dlatego AI nie powinna być analizowana w kategoriach zbawienia czy katastrofy, lecz przez teorię instytucjonalnej niezawodności. James Reason nauczył teorię bezpieczeństwa patrzeć na katastrofę nie jak na błąd jednej osoby, lecz jako na nałożenie się luk w wielu warstwach zabezpieczeń. AI governance powinno działać podobnie: bezpieczeństwo ma wynikać z wielu niezależnych poziomów – modelu, interfejsu, człowieka, audytu, monitoringu, prawa, konkurencji i możliwości odwołania. Wówczas awaria jednej warstwy nie staje się automatycznie katastrofą systemu. To cyfrowy odpowiednik zasady checks and balances.

Sztuczna inteligencja ujawnia zatem bardziej ogólną prawdę o władzy. Niebezpieczna nie jest sama zdolność przewidywania, lecz ta połączona z prawem do decydowania, brakiem kontroli i niemożnością odwołania. Model rekomendujący człowiekowi rozwiązanie jest czymś innym niż system automatycznie pozbawiający go dostępu do świadczenia. Problem powstaje wtedy, gdy epistemiczna przewaga zostaje sprzężona z normatywną mocą decyzji. Da Empoli metaforycznie nazywa to „Zamkiem".

Nauka może określić to precyzyjniej jako koncentrację epistemic authority, decision authority i infrastructural power w jednym układzie. Jeżeli wszystkie trzy zdolności należą do jednego podmiotu – który wie najwięcej o człowieku, podejmuje wobec niego decyzję i kontroluje

infrastrukturę niezbędną do jej wykonania – asymetria staje się szczególnie głęboka. Dobry projekt instytucjonalny powinien więc rozdzielać przynajmniej część tych funkcji. Możemy sformułować zasadę separation of algorithmic powers: jeden podmiot może tworzyć model, ale nie powinien bez zewnętrznej kontroli sam ustalać wszystkich norm jego zastosowania, mierzyć własnej zgodności i rozstrzygać każdej skargi. Jest to cyfrowe rozwinięcie Monteskiuszowskiej intuicji, że koncentracja funkcji zwiększa możliwość arbitralności.

W tym punkcie artykuł powraca do swego pierwszego problemu – asymetrii sprawczości. AI skraca czas analizy, zwiększa skalę decyzji i obniża marginalny koszt klasyfikowania kolejnych ludzi oraz przypadków. Instytucja może podjąć milion decyzji szybciej, niż milion osób jest w stanie je zakwestionować. Powstaje kolejna odsłona arbitrażu temporalnego władzy: decyzja jest automatyczna, odwołanie zaś proceduralne. To być może najważniejsze niebezpieczeństwo algorytmicznego państwa i rynku. Nie musi istnieć zła intencja; wystarczy różnica prędkości. System generuje dziesięć tysięcy rozstrzygnięć na sekundę, podczas gdy człowiek potrzebuje kilku dni na przygotowanie jednej skargi.

Od indywidualnej skargi do systemowej architektury odpowiedzialności

Wówczas prawo do odwołania może istnieć formalnie, pozostając w praktyce niewystarczającym. Konieczne stają się mechanizmy collective contestability: audyty całych klas decyzji, reprezentacja zbiorowa, organizacje konsumenckie, regulatorzy i sądy zdolne kwestionować nie tylko pojedynczy wynik, ale regułę generującą tysiące podobnych rozstrzygnięć. Charles Sabel i Jonathan Zeitlin opisywali eksperymentalne governance oparte na monitorowaniu, benchmarkingach i ciągłym korygowaniu standardów. AI może wymagać właśnie podobnej instytucjonalności: nie tylko indywidualnego prawa skargi, ale systemowego uczenia się z wzorców tych zgłoszeń. Jeżeli sto osób niezależnie sygnalizuje podobny problem, system nie powinien jedynie rozpatrzyć stu spraw, lecz zapytać, czy istnieje wada modelu lub procesu. Procedural justice musi zatem zostać rozszerzona z poziomu jednostkowego na poziom systemowy.

To jest ostateczna odpowiedź na kafkowską metaforę. Człowiek nie powinien być zmuszony przez całe życie chodzić po korytarzach "Zamku", próbując znaleźć osobę odpowiedzialną za wynik. System powinien posiadać adres odpowiedzialności, rejestrować ścieżkę decyzji i sam wykrywać seryjne wzorce błędu. Nie oznacza to powrotu do świata bez automatyzacji; oznacza automatyzację wyposażoną w konstytucję. Jeśli AI ma stać się częścią instytucji, musi przejąć nie tylko ich zdolność do działania, lecz także obowiązek uzasadniania decyzji, przyjmowania skarg,

przyznawania się do błędu i poddawania się zewnętrznej kontroli. W przeciwnym razie otrzymujemy jedynie przyspieszoną administrację bez przyspieszonej odpowiedzialności. "Godzina drapieżnika" nie jest momentem, w którym maszyna staje się drapieżnikiem. Jest godziną, w której ludzka instytucja może użyć maszyny do takiego zwiększenia szybkości, skali i nieprzejrzystości działania, że tradycyjne mechanizmy kontroli przestaną nadążać.

Opór wobec algorytmicznej dominacji nie musi przyjmować postaci odrzucenia AI. Może przybrać znacznie dojrzalszą formę: żądania prawa do zrozumienia funkcji systemu, prawa do zakwestionowania wyniku, prawa do kompetentnego ludzkiego przeglądu, prawa do poprawienia danych, prawa do niezależnego audytu i prawa do funkcjonalnej alternatywy. Kolejny esej Fundacji Dobre Państwo powinien zatem zejść z poziomu infrastruktury i teorii do niewielkiej francuskiej miejscowości.

Lieusaint oraz konflikt burmistrza Michela Bissona z Waze pozwalają zobaczyć wszystkie dotychczasowe problemy w mikroskali: algorytm posiada jasno zdefiniowaną funkcję celu, kierowcy zachowują się racjonalnie indywidualnie, platforma optymalizuje czas przejazdu, a koszty hałasu, ryzyka i ruchu samochodowego zostają przerzucone na lokalną wspólnotę. Czas na studium człowieka negocjującego z funkcją celu. Ważna korekta polega na tym, że historia Lieusaint nie jest wyłącznie opowieścią o bezsilnej jednostce pokonanej przez Waze. Dostępne źródła pokazują, że lokalne interwencje, przebudowa ruchu i późniejsza współpraca przyniosły częściową redukcję problemu. Tym samym przypadek ten staje się ciekawszy: nie opowiada o wszechmocy algorytmu, lecz o możliwości przeprojektowania bodźców, na które algorytm reaguje. Człowiek nie musi pokonać maszyny. Czasem wystarczy zmienić świat, który maszyna mierzy.

> Ostatecznie walka z algorytmiczną dominacją nie polega na próbie pokonania maszyny w jej własnej grze optymalizacji, lecz na odzyskaniu prawa do błędu i sprzeciwu. Prawdziwym zagrożeniem nie jest bowiem inteligentny drapieżnik, lecz świat, w którym szybkość procesora zastępuje sprawiedliwość procedury. Być może najskuteczniejszą formą oporu nie jest zatem walka z funkcją celu, ale odważne przeprojektowanie rzeczywistości, którą ta funkcja mierzy.

Inspiracje

[1] "The Hour of the Predator" Giuliano da Empoli

Urodzony w 1973 roku **Giuliano da Empoli** to włosko-szwajcarski politolog, eseista i pisarz. Wykłada w Sciences Po w Paryżu i był doradcą premiera Włoch. Do jego najważniejszych książek należą nagradzana powieść „Magik z Kremla" oraz esej geopolityczny „The Hour of the Predator".

Giuliano da Empoli *(born August 27, 1973) is an Italian-Swiss author, journalist, and political scientist. He teaches at Sciences Po Paris and is the founding chairman of Volta, a Milan-based think tank dedicated to European policy and political analysis. He previously served as Florence's Deputy Mayor for Culture and as a senior political advisor to Italian Prime Minister Matteo Renzi. Da Empoli's research and essays examine political communications, digital propaganda, algorithms, and power dynamics in modern democracies. In 2022, he achieved widespread international acclaim with his debut novel, Le Mage du Kremlin (The Wizard of the Kremlin), which won the Grand Prix du roman de l'Académie française and was a finalist for the Prix Goncourt. His non-fiction writings, including Les ingénieurs du chaos, analyze how data-driven spin doctors shape contemporary politics.*

Sciences Po / Volta

Literatura uzupełniająca

Książki

Książki autorów przywoływanych i literatura pokrewna:

[1] "Information: A Very Short Introduction"
Luciano Floridi

[2] "An AI Pattern Language"
Madeleine Elish

[3] "AI in Context: The Labor of Integrating New Technologies"
Madeleine Elish

[4] **"The Second Industrial Divide"**
Charles Sabel

[5] **"Assembling Accountability: Algorithmic Impact Assessment for the Public Interest"**
Madeleine Elish

[6] **"Things that make us smart"**
Donald Norman

[7] **"The Design of Everyday Things"**
Donald Norman

[8] **"Don Norman: 3 ways good design makes you happy"**
Donald Norman

[9] **"Noise: A Flaw in Human Judgment"**
Daniel Kahneman

[10] **"The framing of decisions and the psychology of choice"**
Daniel Kahneman

[11] **"Thinking, Fast and Slow"**
Daniel Kahneman

[12] **"SLIP"**
Joseph Weizenbaum

[13] **"ELIZA"**
Joseph Weizenbaum

[14] **"Computer Power and Human Reason: From Judgment to Calculation"**
Joseph Weizenbaum

[15] **"Bayes' theorem"**
Thomas Bayes

[16] **"An Essay towards solving a Problem in the Doctrine of Chances"**
Thomas Bayes

[17] **"Letter from Thomas Bayes to John Canton"**
Thomas Bayes

[18] **"Causality"**
Judea Pearl

[19] **"Causality: Models, Reasoning and Inference"**
Judea Pearl

[20] **"The Book of Why"**
Judea Pearl

[21] **"The Two Fundamental Problems of the Theory of Knowledge"**
Karl Popper

[22] **"Labour and Monopoly Capital"**
Harry Braverman

[23] **"Labor and monopoly capital; the degradation of work in the twentieth century"**
Harry Braverman

[24] **"Against the Current: Essays in the History of Ideas"**
Isaiah Berlin

[25] **"The Fearless Organization: Creating Psychological Safety in the Workplace for Learning, Innovation, and Growth"**
Amy Edmondson

[26] **"Epistemic Injustice: Power and the Ethics of Knowing"**
Miranda Fricker

[27] **"Jak se loví hlupáci: ekonomie manipulace a klamu: nenechte sebou manipulovat"**
George Akerlof

[28] **"Animal Spirits"**
George Akerlof

[29] **"Phishing for Phools"**
George Akerlof

[30] **"The Costs of Accidents"**
Guido Calabresi

[31] **"El futuro del Derecho y Economía. Ensayos para la reflexión y la memoria"**
Guido Calabresi

[32] **"Das Ende der Massenproduktion"**
Charles Sabel

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] **"The human project"**
Luciano Floridi, TPM: The Philosophers' Magazine
2014 | The Philosophers Magazine | DOI: 10.5840/tpm20146676
Źródło

[2] **"Information Ethics: Its Nature and Scope"**
Luciano Floridi
2008 | Information Technology and Moral Philosophy | DOI: 10.1017/cbo9780511498725.004
Źródło

[3] **"From algorithmic accountability to digital governance"**
Jakob Mökander, L. Floridi
2022 | Nature Machine Intelligence | DOI: 10.1038/s42256-022-00504-5
Źródło

[4] **"Machine Unlearning: Its Nature, Scope, and Importance for a “Delete Culture”"**
L. Floridi
2023 | Philosophy & Technology | DOI: 10.1007/s13347-023-00644-5
Źródło

[5] **"Power to the People? Opportunities and Challenges for Participatory AI"**
Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish
2022 | Conference on Equity and Access in Algorithms, Mechanisms, and Optimization | DOI: 10.1145/3551624.3555290
Źródło

[6] **"Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction"**

Madeleine Clare Elish

2019 | Engaging Science Technology and Society | DOI: 10.17351/ests2019.260

Źródło

[7] **""The human body is a black box""**

Mark Sendak, Madeleine Clare Elish, Michael C. Gao, Joseph Futoma, William Ratliff

2020 | DOI: 10.1145/3351095.3372827

Źródło

[8] **"Algorithmic Impact Assessments and Accountability"**

Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish

2021 | ACM Conference on Fairness, Accountability, and Transparency (FAccT) | DOI: 10.1145/3442188.3445935

Źródło

[9] **"Things That Make Us Smart: Defending Human Attributes in the Age of the Machine"**

Rick Robinson, Donald A. Norman

1994 | Design Issues | DOI: 10.2307/1511661

Źródło

[10] **"Designing for Error"**

CLAYTON LEWIS, DONALD A. NORMAN

1986 | User Centered System Design | DOI: 10.1201/9780367807320-20

Źródło

[11] **"Design rules based on analyses of human error"**

Donald A. Norman

1983 | Communications of the ACM | DOI: 10.1145/2163.358092

Źródło

[12] **"Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err"**

Berkeley J. Dietvorst, Joseph P. Simmons, Cade Massey

2014 | Journal of Experimental Psychology General | DOI: 10.1037/xge0000033

Źródło

[13] "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them"

Berkeley J. Dietvorst, Joseph P. Simmons, Cade Massey

2016 | Management Science | DOI: 10.1287/mnsc.2016.2643

Źródło

[14] "People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error"

Berkeley J. Dietvorst, Soaham Bharti

2020 | Psychological Science | DOI: 10.1177/0956797620948841

Źródło

[15] "Consumers Object to Algorithms Making Morally Relevant Tradeoffs Because of Algorithms' Consequentialist Decision Strategies"

Berkeley J. Dietvorst, Daniel M. Bartels

2021 | Journal of Consumer Psychology | DOI: 10.1002/jcpy.1266

Źródło

[16] "Promoting Transparency in Social Science Research"

Edward A. Miguel, Colin F. Camerer, Katherine Casey, J. Cohen, Kevin Esterling, Joseph P. Simmons

2014 | Science | DOI: 10.1126/science.1245317

Źródło

[17] "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant"

Joseph P. Simmons, Leif D. Nelson, Uri Simonsohn

2011 | PsycEXTRA Dataset | DOI: 10.1037/e519702015-014

Źródło

[18] "Computer power and human reason"

Benjamin J. Kuipers, John McCarthy, Joseph Weizenbaum

1976 | ACM SIGART Bulletin | DOI: 10.1145/1045264.1045265

Źródło

[19] "Computer Power and Human Reason: From Judgement to Calculation."

John F. Crowther, Joseph Weizenbaum

1977 | Contemporary Sociology A Journal of Reviews | DOI: 10.2307/2062802

Źródło

[20] "Computer Power and Human Reason: From Judgment to Calculation."

Gerald R. Rising, Joseph Weizenbaum, Walter Freeman

1978 | American Mathematical Monthly | DOI: 10.2307/2321366

Źródło

[21] "FROM 'POLITICS AS A VOCATION'"

Max Weber

1991 | Edinburgh University Press eBooks | DOI: 10.1515/9781474469654-020

Źródło

[22] "CHARISMATIC POLITICS AND THE SYMBOLIC FOUNDATIONS OF POWER"

Max Weber

2008 | Democracy and the Politics of the Extraordinary | DOI: 10.1017/cbo9780511755842.002

Źródło

[23] "XLIII. A letter from the late Reverend Mr. Thomas Bayes, F. R. S. to John Canton, M. A. and F. R. S"

Thomas Bayes

1763 | Philosophical Transactions of the Royal Society of London | DOI: 10.1098/rstl.1763.0044

Źródło

[24] "Interview with Judea Pearl"

Judea Pearl

2022 | Observational Studies | DOI: 10.1353/obs.2022.0007

Źródło

[25] "Understanding how decisions happen in organizations"

James G. March

1996 | Organizational Decision Making | DOI: 10.1017/cbo9780511584169.004

Źródło

[26] "Managing the Risks of Organizational Accidents"

JAMES T. REASON

2016 | Routledge eBooks | DOI: 10.4324/9781315543543

Źródło

[27] "Reflexive Modernization: Politics, Tradition, and Aesthetics in the Modern Social Order."

Alan Sica, Ulrich Beck, Anthony Giddens, Scott M. Lash

1997 | Social Forces | DOI: 10.2307/2580534

Źródło

[28] "The change in concepts needed to account for human behavior in complex dynamic tasks"

Lisanne Bainbridge

1997 | IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans | DOI: 10.1109/3468.568743

Źródło

[29] "A Study of Real-Time Human Decision-Making Using a Plant Simulator"

Lisanne Bainbridge, J. Beishon, J. H. Hemming, M. Splaine

1968 | Journal of the Operational Research Society | DOI: 10.1057/jors.1968.12

Źródło

[30] "WORKSHOP: EXPERT SYSTEMS IN PROCESS CONTROL"

D. Whitfield, P. Andow, L. Bainbridge, K. Duncan, B. Eddershaw

1984 | DOI: 10.1016/b978-0-85295-172-9.50018-4

Źródło

[31] "WORKSHOP: HUMAN RELIABILITY ASSESSMENT TECHNIQUES"

D. Whitfield, P. Andow, L. Bainbridge, K. Duncan, B. Eddershaw

1984 | DOI: 10.1016/b978-0-85295-172-9.50009-3

Źródło

[32] "Scientific Management"

Harry Braverman

2017 | DOI: 10.1002/9781119395485.ch29

Źródło

[33] "Labor and Politics"

Harry Braverman

1958 | Monthly Review | DOI: 10.14452/mr-010-03-04-1958-07-08_9

Źródło

[34] "Designing as reflective conversation with the materials of a design situation"

D.A. Schön

1992 | Knowledge-Based Systems | DOI: 10.1016/0950-7051(92)90020-g

Źródło

[35] "Human Rights and Capabilities"

Amartya Kumar Sen

2005 | Journal of Human Development | DOI: 10.1080/14649880500120491

Źródło

[36] "Elements of a Theory of Human Rights"

Amartya Sen

2017 | Justice and the Capabilities Approach | DOI: 10.4324/9781315251240-6

Źródło

[37] "Methodological fit in management field research"

Amy C. Edmondson, Stacy E. McManus

2007 | Academy of Management Review | DOI: 10.5465/amr.2007.26586086

Źródło

[38] "Learning from Mistakes is Easier Said Than Done: Group and Organizational Influences on the Detection and Correction of Human Error"

Amy C. Edmondson

1996 | The Journal of Applied Behavioral Science | DOI: 10.1177/0021886396321001

Źródło

[39] "Why Hospitals Don't Learn from Failures: Organizational and Psychological Dynamics That Inhibit System Change"

Anita L. Tucker, Amy C. Edmondson

2003 | California Management Review | DOI: 10.2307/41166165

Źródło

[40] "The Local and Variegated Nature of Learning in Organizations: A Group-Level Perspective"

Amy C. Edmondson

2002 | Organization Science | DOI: 10.1287/orsc.13.2.128.530

Źródło

[41] "Humans Judged by Machines: The Rise of Artificial Intelligence in Finance, Insurance, and Real Estate"

Frank A. Pasquale

2021 | Robotics, AI, and Humanity | DOI: 10.1007/978-3-030-54173-6_10

Źródło

[42] "Constituent Power with Checks and Balances"

P. Pettit

2026 | Oxford University Press eBooks | DOI: 10.1093/oxfordhb/9780192873170.013.0026

Źródło

[43] "Interview with Philip Pettit: Neo-Roman Republicanism and Northeast Asia"

Jun-Hyeok Kwak, P. Pettit

2022 | Comparative Political Theory | DOI: 10.1163/26669773-bja10029

Źródło

[44] "ALGORITHMIC CENTRAL PLANNING: BETWEEN EFFICIENCY AND FREEDOM"

Przemyslaw Palka, Rick Brooks, G. Calabresi, Lingxi Chenyang, Hanoch Dagan

2020

Źródło

[45] "Human error: models and management"

JAMES T. REASON

2000 | BMJ | DOI: 10.1136/bmj.320.7237.768

Źródło

[46] "Human Error"

JAMES T. REASON

1990 | Cambridge University Press eBooks | DOI: 10.1017/cbo9781139062367

Źródło

[47] "Errors and violations on the roads: a real distinction?"

JAMES T. REASON, Antony S. R. Manstead, Stephen G. Stradling, James S. Baxter, KAREN CAMPBELL

1990 | Ergonomics | DOI: 10.1080/00140139008925335

Źródło

[48] "Learning from Difference: The New Architecture of Experimentalist Governance in the EU"

Charles Frederick Sabel, Jonathan Zeitlin

2008 | European Law Journal | DOI: 10.1111/j.1468-0386.2008.00415.x

Źródło

[49] "Work and Politics"

Charles Frederick Sabel

1982 | Cambridge University Press eBooks | DOI: 10.1017/cbo9780511571572

Źródło

[50] "HISTORICAL ALTERNATIVES TO MASS PRODUCTION: POLITICS, MARKETS AND TECHNOLOGY IN NINETEEN-THCENTURY INDUSTRIALIZATION"

Charles Frederick Sabel, Jonathan Zeitlin

1985 | Past & Present | DOI: 10.1093/past/108.1.133

Źródło

[51] "Experimentalist Governance in the European Union: Towards a New Architecture"

Charles Frederick Sabel, Jonathan Zeitlin

2010 | bepress Legal Repository

Źródło

[52] "World of Possibilities: Flexibility and Mass Production in Western Industrialization"

Chris Benner, Charles Frederick Sabel, Jonathan Zeitlin

1999 | Economic Geography | DOI: 10.2307/144584

Źródło

Artykuły naukowe autora:

[53] "Architettura Open Source. Verso una progettazione aperta"

Elke Pohl, Assaf Biderman, Michele Bonino, Richard Burdett, Pierre Alain Croset, Giuliano da Empoli

2014

Źródło

[54] "Open Source Architecture"

Ethel Baraona Pohl, Assaf Biderman, Michele Bonino, Ricky Burdett, Pierre Alain Croset, Giuliano da Empoli

2015

Źródło

[55] "L'heure des prédateurs"

Giuliano da Empoli

2026 | Gallimard eBooks | DOI: 10.3917/gall.empol.2026.01

Źródło

[56] "Entretien avec Giuliano Da Empoli autour de L'Heure des prédateurs"

Giuliano da Empoli, Patrick Suter, Fabien Dubosson, Helena Goncalves Leal, Dominik Wüthrich

2026 | Open Access CRIS of the University of Bern | DOI: 10.48620/96620

Źródło

[57] "Riyad, novembre 2024"

Giuliano da Empoli

2025 | Cairn.info

Źródło

[58] "Chicago, novembre 2017"

Giuliano da Empoli

2025 | Cairn.info

Źródło

[59] "Florence, mars 2012"

Giuliano da Empoli

2025 | Cairn.info

Źródło

[60] "Présentation"

Giuliano da Empoli

2025 | Cairn.info

Źródło

[61] "Montréal, septembre 2024"

Giuliano da Empoli

2025 | Cairn.info

Źródło

[62] "Berlin, décembre 2024"

Giuliano da Empoli

2025 | Cairn.info

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: 5dda82b1-850f-4c33-ae55-93a54dc1f4b0