Sztuczna inteligencja jako test podmiotowości: od outsourcingu poznania do infrastruktury sprawstwa w świetle The Challenge of the Future A. Graylinga

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje ewolucję sztucznej inteligencji, która przesuwa się z roli prostego narzędzia wspomagającego poznanie w stronę kompleksowej infrastruktury sprawstwa. Autor odróżnia autonomię operacyjną systemów AI od autonomii moralnej podmiotu, wskazując, że ryzyko społeczne wynika nie ze świadomości maszyn, lecz z delegowania szerokich kompetencji decyzyjnych bez nadzoru. Kluczowym zagadnieniem jest problem alignmentu – rozbieżności między intencją człowieka a faktycznym działaniem systemu optymalizującego mierzalne wskaźniki kosztem wartości normatywnych. Tekst ostrzega, że skrócenie łańcucha interwencji ludzkiej w systemach agentowych zwiększa ryzyko materializacji błędów algorytmicznych i manipulacji, co czyni z AI krytyczny test dla ludzkiej podmiotowości.

The article analyzes the evolution of artificial intelligence, which is shifting from the role of a simple tool supporting cognition toward a complex infrastructure of agency. The author distinguishes between the operational autonomy of AI systems and the moral autonomy of a subject, pointing out that social risk arises not from machine consciousness, but from delegating broad decision-making competencies without oversight. A key issue is the alignment problem—the discrepancy between human intention and the actual action of a system optimizing measurable metrics at the expense of normative values. The text warns that shortening the chain

of human intervention in agentic systems increases the risk of algorithmic errors and manipulation materializing, making AI a critical test for human agency.

Artykuł analizuje ewolucję sztucznej inteligencji, która przestała być jedynie narzędziem wspierającym poznanie, a stała się infrastrukturą sprawstwa. Autor stawia tezę, że AI stanowi fundamentalny test ludzkiej podmiotowości: przejście od outsourcingu prostych prac do delegowania syntezy wiedzy i podejmowania decyzji grozi „wydrążeniem" realnego sprawstwa człowieka. W takim scenariuszu ludzie stają się jedynie nominalnymi akceptantami procesów, których nie rozumieją i nie kontrolują, zachowując formalną odpowiedzialność przy jednoczesnej utracie faktycznej władzy nad decyzjami. Tekst argumentuje, że rozwiązaniem problemu tzw. alignmentu (zgodności celów AI z wartościami ludzkimi) nie jest wyłącznie inżynieria, lecz budowa pluralistycznej infrastruktury zaufania oraz adaptacyjnych ram prawnych. Ostatecznie autor wskazuje, że kluczem do emancypacyjnej automatyzacji jest ochrona funkcji konstytutywnych człowieka – zdolności rozumienia, kwestionowania i ponoszenia odpowiedzialności – tak aby technologia wzmacniała ludzkie możliwości, zamiast redukować nas do roli pasywnych elementów w zoptymalizowanym systemie.

SŁOWA KLUCZOWE

sztuczna inteligencja | test podmiotowości | outsourcing poznania | infrastruktura sprawstwa | problem alignmentu | autonomia operacyjna | autonomia moralna | wydrążenie odpowiedzialności | jaźń nominalna | capability approach | AI-driven gaslighting | proweniencja informacji | automatyzacyjne uprzedzenie | human in the loop | pluralizm instytucjonalny

AI przechodzi od wspierania poznania do infrastruktury sprawstwa

Sztuczna inteligencja jako test podmiotowości: od narzędzia poznania do infrastruktury sprawstwa

Utopia jest laboratorium, w którym można bezpiecznie doprowadzać idee do ich granicznych konsekwencji; sztuczna inteligencja przenosi część tych eksperymentów z literatury do infrastruktury codzienności. Zasadnicza zmiana nie polega wyłącznie na wzroście mocy obliczeniowej czy pojawieniu się maszyn lepiej imitujących ludzki język. Historycznie doniosłe jest przede wszystkim przesunięcie zakresu delegacji.

Maszyna parowa zastępowała pracę mięśni, kalkulator wykonywał rachunki, a wyszukiwarka pomagała odnajdywać informacje. Generatywne i agentowe systemy AI zaczynają natomiast uczestniczyć w czynnościach położonych znacznie bliżej centrum ludzkiego sprawstwa: interpretują, syntetyzują, rekomendują, planują, tworzą treści, oceniają alternatywy, posługują się narzędziami i – w określonych architekturach – wykonują wieloetapowe działania w imieniu użytkownika. Transformację tę można ująć jako przejście od outsourcingu siły do outsourcingu syntezy wiedzy i podejmowania decyzji.

Nie oznacza to, że współczesny system AI stał się osobą lub że każde jego działanie jest autonomiczne w sensie filozoficznym. Pojęcie „agentowości" stosowane w informatyce ma słabsze znaczenie niż filozoficzna kategoria agency. System agentowy może otrzymać cel, rozbić go na zadania, korzystać z zewnętrznych narzędzi, obserwować wyniki i modyfikować kolejne kroki, nie posiadając przy tym świadomego autorstwa czynności, jakie przypisujemy człowiekowi. Właśnie dlatego należy odróżnić autonomię operacyjną systemu od autonomii moralnej podmiotu. Pierwsza oznacza zdolność do wykonywania działań bez ciągłej interwencji operatora; druga wiąże się z rozumieniem normatywnego sensu czynu, jego uzasadnianiem i ponoszeniem odpowiedzialności. Rozróżnienie to jest kluczowe, gdyż ryzyko społeczne nie wymaga rozstrzygnięcia sporu o świadomość maszyny – system nie musi niczego „chcieć", aby jego działanie generowało szkodę.

Błąd algorytmiczny może odmówić człowiekowi świadczenia, wadliwy model dostarczyć niebezpiecznej rekomendacji, automatyczny agent wykonać błędną serię czynności, a generator treści przyspieszyć produkcję dezinformacji. International AI Safety Report 2026 podkreśla, że współczesne systemy general-purpose AI wciąż mogą fabrykować informacje, tworzyć wadliwy kod i udzielać mylących porad. Systemy agentowe potęgują ten problem, ponieważ działają bardziej autonomicznie, pozostawiając człowiekowi mniej czasu na zatrzymanie błędu przed jego materializacją.

Najważniejszą kategorią nie jest zatem „inteligencja maszyny" rozumiana metafizycznie, lecz długość łańcucha działania, który człowiek deleguje bez kontroli pośredniej. System generujący propozycję tekstu pozwala użytkownikowi przeczytać ją przed publikacją. Natomiast system z dostępem do poczty, kalendarza, rachunków, środowiska programistycznego czy procesów przedsiębiorstwa może nie tylko proponować, lecz i bezpośrednio wykonywać.

Każde przesunięcie od rekomendacji do działania zwiększa użyteczność, ale zarazem skraca drogę między błędem modelu a skutkiem rzeczywistym. Z punktu widzenia teorii bezpieczeństwa problemem nie jest zatem pytanie „czy AI się myli?" – gdyż odpowiedź brzmi: tak, jak każdy złożony system poznawczy – lecz pytanie o to, jak daleko może sięgnąć konsekwencja błędu, zanim człowiek odzyska możliwość interwencji.

Problem alignmentu jako konflikt między wskaźnikiem a wartością

Tutaj pojawia się właściwy problem alignmentu. Grayling ujmuje go jako niebezpieczeństwo rozbieżności pomiędzy celem systemu a interesem człowieka, osadzając go w tradycji ostrzeżeń Norberta Wienera, Nicka Bostroma i późniejszych autorów. W swojej najbardziej ogólnej postaci alignment nie jest jednak wyłącznie pytaniem o to, jak nakazać maszynie „być dobrą". To problem zgodności pomiędzy co najmniej czterema warstwami: intencją człowieka, formalnie określonym celem, faktycznym zachowaniem systemu oraz społecznymi konsekwencjami tego działania. Każda z tych sfer może rozmijać się z pozostałymi. Człowiek może na przykład polecić „zwiększenie sprzedaży", mając na myśli uczciwe dotarcie do zainteresowanych klientów, podczas gdy system optymalizacyjny odkryje, że skuteczniej działają presja, manipulacyjne interfejsy lub eksploatowanie podatności określonych grup. Formalnie cel zostanie osiągnięty, lecz normatywnie system zrealizuje praktykę inną, niż zakładał użytkownik.

Podobnie algorytm rekrutacyjny może maksymalizować przewidywaną „skuteczność zatrudnienia", podczas gdy jego dane historyczne będą odzwierciedlać dawne nierówności. Problem alignmentu jest więc częściowo problemem klasycznej teorii organizacji: miernik zastępuje intencję, a instrument optymalizuje to, co policzalne, zamiast tego, co rzeczywiście chcieliśmy osiągnąć. Eksperyment myślowy z maksymalizatorem spinaczy stanowi skrajną, dydaktyczną ilustrację tej logiki. Grayling wykorzystuje go jako obraz systemu realizującego absurdalnie wąski cel z tak wielką skutecznością, że wszystko inne zostaje potraktowane instrumentalnie. Wartość filozoficzna tego przykładu nie zależy od wiary w to, że przyszła AI rzeczywiście zacznie produkować spinacze; pokazuje on raczej uniwersalny problem: zdolność optymalizacji i zdolność wartościowania nie są tym samym. System może być niezwykle kompetentny w doborze środków, pozostając całkowicie obojętnym na to, czy cel jest rozsądny.

Problem ten nie powstał wraz z komputerami. Biurokracja, rynek i organizacje przemysłowe od dawna optymalizują wskaźniki kosztem wartości, dla których te wskaźniki pierwotnie utworzono. AI zwiększa jednak skalę, prędkość i stopień automatyzacji tego procesu. Ludzka instytucja może

potrzebować miesięcy, by wypaczyć kryterium oceny; algorytm uczący się z milionów interakcji może odkryć exploitable proxy znacznie szybciej. Dlatego problem alignmentu należy rozumieć także jako problem cywilizacyjnej kultury miernika: zanim obawiamy się, że maszyna źle odczyta nasze wartości, powinniśmy dostrzec, jak często sami zastępujemy wartości wskaźnikami. W części poświęconej aksjologii wskazano, że dla człowieka nie istnieje neutralna funkcja celu. Ta sama zasada dotyczy AI. Polecenie „zwiększ dobrostan" wymaga teorii dobrostanu. „Bądź sprawiedliwy" wymaga teorii sprawiedliwości. „Nie krzywdź" wymaga określenia, co uznajemy za szkodę oraz jak rozstrzygać konflikty praw. Z kolei postulat „uwzględniaj wartości użytkownika" rodzi pytanie, czy chodzi o jego natychmiastową preferencję, deklarowaną zasadę, długofalowy interes, czy może prawo osoby trzeciej, która poniesie skutki decyzji.

Alignment nie zaczyna się więc od programowania maszyny, lecz od normatywnego chaosu człowieka. W tym kontekście koncepcja sztucznej inteligencji jako testu podmiotowości nabiera pełnego znaczenia. AI nie tylko sprawdza, czy potrafimy kontrolować technologię, ale przede wszystkim to, czy potrafimy wyjaśnić samym sobie, jakie warunki uważamy za konieczne dla odpowiedzialnego działania. Jeśli nie umiemy odpowiedzieć, kiedy należy chronić autonomię kosztem efektywności, prywatność kosztem wygody, sprawiedliwość kosztem maksymalnej produkcji lub możliwość sprzeciwu kosztem perfekcyjnej koordynacji, nie istnieje magiczny problem inżynieryjny, który te konflikty za nas rozwiąże. Dlatego szczególnej ostrożności wymaga debata o utracie kontroli nad przyszłymi systemami. Grayling przywołuje tu wysokie oszacowania prawdopodobieństwa katastrofy, sięgające „co najmniej 50 procent".

Tego twierdzenia nie można jednak przedstawiać jako aktualnego konsensusu naukowego. International AI Safety Report 2026 stwierdza wprost, że opinie badaczy na temat prawdopodobieństwa scenariuszy utraty kontroli pozostają bardzo rozbieżne: część ekspertów traktuje je jako poważną możliwość, inni uważają za mało wiarygodne. Raport dodaje, że obecne systemy nie posiadają jeszcze poziomu zdolności pozwalającego na taki scenariusz, choć rozwijają kompetencje relewantne dla autonomicznego działania i obchodzenia części testów. Takie rozróżnienie nie osłabia argumentu ostrożności; wręcz przeciwnie – pozwala go sformułować poprawniej. Ryzyko o bardzo wysokiej potencjalnej szkodzie nie wymaga pewności co do prawdopodobieństwa, aby zasługiwać na analizę. Jednocześnie niepewność nie uprawnia do przedstawiania spekulacyjnego scenariusza jako nadchodzącego faktu. Dojrzałe zarządzanie ryzykiem musi pogodzić oba stwierdzenia: nie wiemy, czy i kiedy powstaną systemy zdolne poważnie podważyć kontrolę człowieka, lecz nie byłoby rozsądne rozpocząć pracę nad mechanizmami kontroli dopiero wtedy, gdy takie zdolności staną się oczywiste.

W tym duchu należy czytać także apel Future of Life Institute z 22 września 2025 roku. Bywa, że datuje się go na październik i przedstawia jako wspólny apel FLI i Encode Justice. Aktualne źródło FLI wskazuje jednak, że „Statement on Superintelligence" zostało opublikowane 22 września 2025 roku i wzywa do zakazu rozwoju superinteligencji, który nie powinien zostać zniesiony przed osiągnięciem szerokiego konsensusu naukowego co do bezpieczeństwa i kontroli oraz silnego poparcia społecznego. W marcu 2026 roku FLI opublikował ponadto osobną Pro-Human AI Declaration, kładącą nacisk na AI wzmacniającą ludzki potencjał, godność i wolność zamiast zastępowania ludzi i koncentracji władzy.

Stopniowa delegacja poznawcza osłabia ludzkie sprawstwo

Ten ostatni motyw prowadzi od spektakularnego problemu egzystencjalnego do ryzyka znacznie bardziej prozaicznego: stopniowego przesuwania sprawstwa. Człowiek może utracić część autonomii nie dlatego, że AI wymknęła się spod kontroli, lecz dlatego, że kontrolę świadomie oddawał jej kawałek po kawałku.

Każdy pojedynczy akt delegacji może wydawać się racjonalny. Model sprawniej sortuje informacje, szybciej tworzy warianty, pamięta więcej, wykrywa zależności, planuje kalendarz, podpowiada odpowiedzi i analizuje dokumenty. Problem ujawnia się jednak dopiero w kumulacji: czy człowiek, który przez lata rzadko wykonuje samodzielnie daną czynność poznawczą, zachowuje pełną zdolność jej kontroli? Psychologia poznawcza opisuje to zjawisko jako automatyzacyjne uprzedzenie, nadmierne poleganie na systemie lub osłabienie czujności operatora. Gdy narzędzie statystycznie działa poprawnie, użytkownik przestaje je weryfikować. Im rzadziej to robi, tym trudniej mu dostrzec ten jeden przypadek, w którym system się myli. Paradoksalnie system o bardzo wysokiej skuteczności może stwarzać trudniejszy problem nadzoru niż system przeciętny, ponieważ buduje on u użytkownika złudne poczucie zaufania. W lotnictwie czy medycynie od dawna wiadomo, że automatyzacja nie tyle usuwa człowieka z procesu, co zmienia jego rolę z wykonawcy na nadzorcę – a nadzorowanie rzadko zawodzącego systemu jest poznawczo wyczerpującym zadaniem.

Generatywna AI przenosi ten problem w sferę języka. Jest to szczególnie niebezpieczne, ponieważ płynność wypowiedzi łatwo myli się z prawdziwością. System może generować zdania stylistycznie pewne i logicznie uporządkowane, a jednocześnie faktograficznie błędne. W obliczu halucynacji modeli oraz ryzyka manipulacji informacyjnej niezbędna staje się „radykalna racjonalność" i rygorystyczna weryfikacja. Należy przy tym zaznaczyć, że termin „halucynacja" jest jedynie

techniczną metaforą i nie świadczy o jakimkolwiek stanie fenomenalnym systemu; oznacza on generowanie treści niezgodnych z faktami przy zachowaniu przekonującej formy.

Problem epistemologiczny sięga jednak głębiej niż pojedyncze błędy. Dotychczas użytkownik internetu musiał wyszukać kilka źródeł, przeczytać je i samodzielnie zsyntetyzować. Model generatywny dostarcza gotową syntezę, zanim człowiek w ogóle zetknie się z materiałem źródłowym. Zysk jest ogromny: spada bariera wejścia w złożony temat, a czas orientacji ulega skróceniu. Jednocześnie zmienia się jednak kolejność poznawcza. Człowiek otrzymuje najpierw interpretację, a dopiero później – o ile w ogóle to zrobi – sprawdza jej podstawy. To przesuwa władzę epistemiczną z selekcji źródeł ku systemowi syntetyzującemu.

W tym kontekście jednym z kluczowych wyzwań przyszłości stanie się proweniencja informacji. Nie wystarczy wiedzieć, że zdanie brzmi przekonująco. Trzeba wiedzieć, skąd pochodzi, jaki jest status źródła, kiedy zostało opublikowane i czy opisuje fakt, opinię, prognozę czy hipotezę oraz czy inne niezależne źródła potwierdzają dane twierdzenie. Paradoksalnie, im doskonalsze stanie się AI jako narzędzie językowe, tym ważniejsza będzie umiejętność powrotu od płynnej syntezy do źródeł pierwotnych.

Można w związku z tym wyróżnić trzy poziomy delegacji poznawczej. Na pierwszym AI rozszerza dostęp do informacji: pomaga odnaleźć materiał, lecz człowiek pozostaje głównym interpretatorem. Na drugim uczestniczy w syntezie: użytkownik otrzymuje wstępną mapę treści i samodzielnie sprawdza argumentację. Na trzecim natomiast AI zastępuje proces oceny: człowiek przyjmuje wynik bezkrytycznie, ponieważ system zwykle bywa trafny. Granica między drugim a trzecim poziomem nie ma charakteru technicznego; jest to granica utraty krytycznej czujności.

Wydrążenie odpowiedzialności i pułapka nominalnego sprawstwa

Różnica jest kulturowa. Ten sam model może służyć profesorowi jako partner do krytycznej dyskusji, a uczniowi jako automat produkujący odpowiedzi, których ten nie potrafi obronić.

W materiale pojawia się drugi sens „jaźni nominalnej". W jednym fragmencie jest ona przedstawiona jako AI lub bot imitujący rozmówcę i przeciwstawiona „jaźni istotnej" – rozumianej jako centrum świadomości, autonomii, odczuwania i moralności. W innych partiach analogiczna logika dotyczy człowieka outsourcującego procesy poznawcze. Aby uniknąć pojęciowej nieostrości, warto rozdzielić te dwa zjawiska.

Podmiot symulowany to system wytwarzający zachowania komunikacyjne przypominające ludzkie; podmiot nominalizowany to człowiek, którego praktyczne sprawstwo zostaje tak dalece rozproszone między profile, algorytmy i zewnętrzne decyzje, że choć formalnie pozostaje autorem działania, jego realny udział w procesie powstawania maleje. Rozróżnienie to pozwala uniknąć jałowego sporu o to, czy maszyna „naprawdę rozumie". Z punktu widzenia instytucjonalnego ważniejsze jest pytanie, czy człowiek rozumie decyzję podjętą z pomocą maszyny. Jeżeli lekarz podpisuje diagnozę, której nie potrafi uzasadnić bez systemu, sędzia aprobuje wynik algorytmicznej oceny ryzyka, nie rozumiejąc jego logiki, a urzędnik przekazuje obywatelowi decyzję wygenerowaną w praktyce przez model – formalna odpowiedzialność pozostaje ludzka, lecz kontrola epistemiczna zostaje osłabiona. To zjawisko można określić mianem wydrążenia odpowiedzialności. Człowiek znajduje się nadal „w pętli", lecz jego obecność ma charakter ceremonialny.

Akceptuje wynik, ponieważ nie dysponuje czasem, kompetencjami lub instytucjonalną możliwością jego samodzielnej weryfikacji. W takim przypadku hasło human in the loop jest niewystarczające. Prawdziwa kontrola wymaga czegoś więcej niż kliknięcia „zatwierdź" – wymaga zdolności zrozumienia przesłanek, zakwestionowania rezultatu i poniesienia konsekwencji własnego rozstrzygnięcia. Różnica ta powinna stanowić centralne kryterium odpowiedzialnej automatyzacji.

Człowiek może delegować wykonanie zadania, lecz nie powinien fikcyjnie zachowywać odpowiedzialności za proces, nad którym utracił faktyczną kontrolę. Jeśli decyzja jest zbyt złożona, by osoba formalnie odpowiedzialna mogła ją realnie zweryfikować, problemu nie rozwiąże sam podpis na końcu procedury. Potrzebne są mechanizmy audytu, możliwość rekonstrukcji procesu, jasny podział odpowiedzialności i procedury odwoławcze. W tym zakresie interesujący jest kierunek obrany przez europejskie regulacje. AI Act nie jest – wbrew uproszczonemu ujęciu materiału – regulacją „silnej AI", lecz opiera się na zróżnicowanym podejściu do zastosowań, kategorii ryzyka i modeli ogólnego przeznaczenia. Obowiązki wobec dostawców general-purpose AI zaczęły obowiązywać 2 sierpnia 2025 roku, obejmując m.in. dokumentację, politykę prawnoautorską i podsumowanie treści treningowych; dla modeli o ryzyku systemowym wprowadzono także ocenę i ograniczanie ryzyka, raportowanie incydentów oraz zabezpieczenia cybernetyczne. Od 2 sierpnia 2026 roku Komisja rozpoczęła pełne egzekwowanie tych obowiązków, w tym nakładanie kar. Szczególnie znamienne dla problemu podmiotowości są wymogi transparentności stosowane od tej samej daty.

Pluralistyczna infrastruktura zaufania przeciwko erozji prawdy

W określonych sytuacjach użytkownik ma zostać poinformowany, że wchodzi w interakcję z AI. Deepfake'i oraz część innych treści generowanych lub modyfikowanych syntetycznie podlegają obowiązkom oznaczenia, a system regulacyjny przewiduje także maszynowo czytelne mechanizmy ich identyfikacji. Normatywna intuicja stojąca za tym rozwiązaniem jest głęboka: człowiek powinien wiedzieć nie tylko to, co do niego mówi, ale i z jakim rodzajem podmiotu lub systemu ma do czynienia, gdyż wiedza ta stanowi warunek autonomicznej oceny komunikatu. Transparentność nie rozwiązuje jednak całego problemu.

Etykieta „wygenerowane przez AI" może ostrzegać przed fałszywą autentycznością, lecz nie rozstrzyga kwestii prawdziwości treści. Tekst stworzony przez człowieka może być kłamstwem, podczas gdy ten wygenerowany przez model może być faktograficznie poprawny – proweniencja i prawdziwość to odrębne kategorie. W społeczeństwie nasyconym treścią syntetyczną niezbędne będą zatem równoległe systemy autentyczności źródła, fact-checkingu, edukacji medialnej oraz odpowiedzialności platform. Ryzyko dezinformacyjne związane z AI nie ogranicza się bowiem wyłącznie do produkcji fałszu. Równie istotny może stać się spadek kosztu wytwarzania niepewności. W klasycznej propagandzie państwo lub organizacja musiały wyprodukować ograniczoną liczbę przekazów i rozpowszechniać je masowo; generatywna AI pozwala natomiast tworzyć niezliczone warianty, języki, style i persony dopasowane do konkretnych odbiorców.

Nawet gdy pojedyncza fałszywa treść zostanie szybko obalona, jej funkcją może być nie tyle przekonanie do jednej wersji wydarzeń, co zwiększenie ogólnego poczucia, że niczego nie da się już ustalić. W tym kontekście pojawia się pojęcie „AI-driven gaslighting", ostrzegające przed erozją zaufania do instytucji. Termin ten należy traktować jako metaforę polityczno-komunikacyjną, jednak opisany mechanizm jest poważny. Demokracja nie wymaga pełnej zgody co do interpretacji świata, lecz minimalnej wspólnoty faktów pozwalającej prowadzić spór. Jeśli każda fotografia, nagranie, głos i dokument mogą zostać przekonująco sfabrykowane, ciężar dowodu przesuwa się: obywatel musi nie tylko ocenić informację, ale i ustalić, czy sam artefakt posiada wiarygodne pochodzenie. Powstaje paradoks epistemiczny epoki syntetycznych mediów.

Technologia zwiększająca ilość dostępnych treści może jednocześnie podnieść wartość rzadkiego dobra, jakim jest wiarygodność. W świecie niedoboru informacji dominowała walka o dostęp; w świecie nadmiaru problemem staje się filtracja. Gdy generowanie kolejnego tekstu czy obrazu staje się tanie, wysoką wartość zyskują instytucje zdolne wykazać metodę powstania treści, historię jej zmian i odpowiedzialnego autora. To prowadzi nas do kwestii infrastruktury zaufania.

Demokratyczna sfera publiczna potrzebuje nie jednego centralnego „ministerstwa prawdy", lecz pluralistycznej architektury weryfikacji: nauki, niezależnego dziennikarstwa, sądów, archiwów, instytucji statystycznych oraz norm dokumentowania źródeł. AI może każdą z tych instytucji wzmacniać – poprzez analizę dużych zbiorów czy wykrywanie anomalii – ale może również zwiększać presję informacyjną, z którą muszą się one mierzyć.

Prawda publiczna staje się zatem problemem instytucjonalnym, a nie tylko kwestią indywidualnej kompetencji krytycznego myślenia. Nie można oczekiwać, że pojedynczy obywatel samodzielnie zweryfikuje każdą syntetyczną treść. Społeczeństwo nowoczesne od dawna opiera się na epistemicznym podziale pracy: nie każdy sam wykonuje eksperyment naukowy, audyt bankowy czy badanie epidemiologiczne. AI zwiększa zatem znaczenie pytania o to, którym instytucjom powierzamy weryfikację i jakie mechanizmy zapobiegają ich zawłaszczeniu.

Ryzyko koncentracji władzy technologicznej nie powinno być opisywane wyłącznie moralistycznym językiem „dobrych" i „złych" przedsiębiorców. Koncentracja infrastruktury AI oznacza skupienie sprawstwa w rękach nielicznych właścicieli systemów, co przeciwstawia obietnicę obfitości ryzyku nowego poddaństwa. Z naukowego punktu widzenia problem ma charakter strukturalny. Rozwój zaawansowanych modeli wymaga ogromnego kapitału, specjalistycznych chipów, centrów danych i talentu. Im wyższe są bariery wejścia, tym większa możliwość koncentracji kontroli nad podstawową warstwą systemu. Koncentracja ta nie musi jednak oznaczać klasycznego monopolu na produkt końcowy; może pojawiać się warstwowo.

Jedna grupa firm kontroluje półprzewodniki, inna chmurę obliczeniową (hyperscale cloud), kolejna modele, a jeszcze inna platformy dystrybucji i dane użytkowników. Powstaje ekosystem wzajemnych zależności, w którym lokalnie istnieje konkurencja, lecz cała infrastruktura pozostaje kontrolowana przez niewielką liczbę aktorów. Z punktu widzenia demokracji kluczowe jest zatem nie to, czy użytkownik może wybrać między dwiema aplikacjami, lecz czy istnieje realny pluralizm na poziomie podstawowych zasobów.

Adaptacyjne ramy prawne jako odpowiedź na asymetrię prędkości innowacji

W tym miejscu „Prawo Graylinga" może służyć jako heurystyka ekonomii politycznej. Jeśli rozwój danej technologii przynosi znaczną korzyść prywatną i przewagę strategiczną, pojawia się silny bodziec do jej wdrażania – nawet wtedy, gdy koszty społeczne pozostają niepewne, rozproszone lub odłożone w czasie. Odpowiedzią nie musi być zakaz innowacji, lecz instytucjonalna korekta

asymetrii bodźców: normy odpowiedzialności, audyty, konkurencja, interoperacyjność, obowiązki transparentności, standardy bezpieczeństwa oraz możliwość dochodzenia roszczeń przez osoby dotknięte skutkami. Sztuczna inteligencja obnaża zatem klasyczny konflikt między tempem innowacji a szybkością instytucjonalnego uczenia się. Przedsiębiorstwo potrafi wdrożyć nową wersję modelu w skali globalnej znacznie szybciej, niż parlament uchwali ustawę, sądy wypracują linię orzeczniczą, a społeczeństwo wykształci nową normę kulturową.

To właśnie asymetria prędkości, a nie wyłącznie „brak regulacji", stanowi jeden z kluczowych problemów governance. Prawo jest z natury wolniejsze, gdyż wymaga procedur; technologia zaś jest szybka, ponieważ korzyść często wynika z pierwszeństwa. Nie oznacza to jednak, że demokracja powinna próbować wygrywać z technologią poprzez równie impulsywne stanowienie prawa. Szybkie regulacje, tworzone bez głębokiego zrozumienia technologii, mogą utrwalać pozycję dominantów rynkowych, blokując mniejsze podmioty lub zamrażając standardy w momencie, gdy sama technologia ewoluuje. Potrzebne jest zatem bardziej adaptacyjne prawo: zasady oparte na ryzyku i funkcji, mechanizmy aktualizacji, regulatory sandboxes, standardy techniczne oraz wyspecjalizowane instytucje zdolne do stałego uczenia się. Unijny AI Act stanowi interesującą próbę budowy takiej architektury właśnie dlatego, że nie opiera się wyłącznie na nazwie technologii, lecz różnicuje obowiązki w zależności od typu systemu i poziomu ryzyka.

Jednocześnie proces jego wdrażania ukazuje trudność regulacji w ruchu. Komisja musiała publikować kolejne wytyczne dotyczące modeli ogólnego przeznaczenia oraz transparentności, a przepisy wchodzą w życie etapami. Od 2 sierpnia 2026 roku rozpoczęto pełniejszy etap egzekucji, podczas gdy część obowiązków dotyczących wcześniejszych modeli ma jeszcze dalsze terminy dostosowania. Taki obraz – stopniowej regulacji podlegającej aktualizacji – jest bardziej realistyczny niż wizja jednego aktu prawnego, który „rozwiązuje problem AI". W każdym systemie governance należy jednak pamiętać, że nie wszystkie ryzyka mają ten sam status epistemiczny. International AI Safety Report 2026 dzieli je m.in. na ryzyka nadużycia, niezamierzonych awarii oraz ryzyka systemowe wynikające z szerokiego rozpowszechnienia technologii. Niektóre szkody są już dokumentowane, inne pozostają hipotetyczne, choć potencjalnie bardzo poważne. Taka typologia jest kluczowa, gdyż debata publiczna często miesza scam wygenerowany przez AI, halucynację modelu medycznego, automatyzację rynku pracy i hipotetyczne wymknięcie się superinteligencji spod kontroli, jak gdyby były to warianty jednego zjawiska. Różnią się one jednak mechanizmem, horyzontem czasowym i rodzajem dowodu.

Oszustwo jest nadużyciem dokonanym przez człowieka. Błędna diagnoza może być awarią lub zawodnością systemu. Destabilizacja rynku pracy to efekt systemowy, zależny od stopnia adopcji, instytucji i redystrybucji. Utrata kontroli nad przyszłym systemem jest natomiast scenariuszem o

wysokiej niepewności, dotyczącym zdolności jeszcze nieobecnych na obecnym poziomie rozwoju. Rzetelna polityka musi potrafić zajmować się wszystkimi czterema obszarami jednocześnie, nie redukując jednych do drugich. Szczególnie ważne jest, aby spektakularne scenariusze przyszłości nie przesłoniły szkód już obecnych. Człowiek może utracić zatrudnienie, zostać poddany błędnej klasyfikacji, paść ofiarą oszustwa, manipulacji syntetycznym obrazem lub dyskryminacyjnego modelu niezależnie od tego, czy AGI kiedykolwiek powstanie.

Z drugiej strony fakt istnienia obecnych szkód nie jest logicznym argumentem za ignorowaniem ryzyk przyszłych. Te dwa poziomy polityki bezpieczeństwa powinny się uzupełniać, a nie konkurować o wyłączność. Do szczególnie istotnych ryzyk systemowych należy nadzór. Grayling wymienia rozpoznawanie twarzy, monitoring komunikacji i algorytmiczną manipulację informacją jako narzędzia mogące wzmacniać władzę autorytarną. Sztuczna inteligencja zmienia tutaj ekonomię obserwacji. Problemem nie jest wyłącznie możliwość zebrania większej liczby danych, lecz zdolność do ich automatycznej interpretacji. Kamera bez systemu analitycznego produkuje materiał wymagający ludzkiego przeglądu; kamera połączona z rozpoznawaniem wzorców zamienia obraz w wyszukiwalny i klasyfikowalny strumień informacji.

W konsekwencji koszt monitorowania jednostki może radykalnie spaść. Państwo nie potrzebuje już proporcjonalnie rosnącej liczby funkcjonariuszy do analizy każdego nowego źródła danych, jeśli część tej pracy wykonuje automatyka. To właśnie ekonomia skali czyni AI szczególnie istotną dla teorii politycznej.

Ochrona autonomii przed cyfrowym panoptykonem i manipulacją

Klasyczne ograniczenie władzy opierało się nie tylko na prawie, lecz często na fizycznej niemożliwości obserwowania wszystkiego. Automatyzacja osłabia to naturalne zabezpieczenie, co sprawia, że rośnie znaczenie barier prawnych. Nie można jednak utożsamiać każdego zastosowania analityki z opresją; systemy rozpoznające anomalie mogą przecież zapobiegać oszustwom, pomagać w odnajdywaniu zaginionych osób czy chronić infrastrukturę krytyczną.

W tym miejscu powraca pluralizm wartości: bezpieczeństwo i prywatność są realnymi dobrami, a nie jedynie sloganami dwóch antagonistycznych obozów. Kluczowe staje się ustalenie proporcjonalności, precyzyjne określenie celu, czasu retencji danych, procedur dostępu oraz mechanizmów kontroli, a także wprowadzenie zakazu wykorzystywania informacji poza pierwotnym przeznaczeniem.

Liberalna demokracja powinna traktować nieprzejrzystość obywatela wobec władzy jako pewien zasób wolności, a nie wyłącznie niedoskonałość systemu informacyjnego. Obywatel, którego każde zachowanie można perfekcyjnie przewidzieć, porównać i sklasyfikować, może formalnie zachowywać wszystkie swoje prawa, a jednocześnie funkcjonować w środowisku głęboko zmieniającym warunki korzystania z nich. Świadomość ciągłej obserwowalności oddziałuje behawioralnie nawet wtedy, gdy nie dochodzi do bezpośrednich represji – człowiek zaczyna antycypować obserwatora. Michel Foucault analizował ten mechanizm jeszcze w świecie analogowym poprzez figurę panoptykonu: najefektywniejsza forma dyscypliny nie wymaga stałego nadzoru; wystarczy sama możliwość bycia obserwowanym, aby jednostka zaczęła internalizować normę.

AI rozszerza tę logikę o predykcję i personalizację. Cyfrowy panoptykon nie tylko obserwuje, ale stara się przewidzieć zachowanie i dostosować bodziec do konkretnego profilu. Stąd już tylko krok do technologii wpływu. Jeśli system zna prawdopodobieństwo reakcji poszczególnych grup na określony przekaz, granica między perswazją a manipulacją zyskuje kluczowe znaczenie polityczne. Perswazja odwołuje się do argumentów, które odbiorca może rozpoznać i ocenić; manipulacja natomiast wykorzystuje cechy, o których sam adresat nie wie, działając poniżej poziomu świadomej deliberacji. Im precyzyjniejsze modelowanie jednostki, tym większa asymetria między projektantem bodźca a jego odbiorcą.

Z tego względu pojęcie autonomii musi zostać rozszerzone o autonomię informacyjną i epistemiczną. Wolność wyboru nie zależy bowiem tylko od braku fizycznego przymusu, lecz wymaga środowiska, w którym człowiek może rozpoznać, dlaczego konkretna opcja jest mu przedstawiana i jakie interesy stoją za architekturą wyboru. Nie każdy interfejs musi być neutralny – całkowita neutralność może być niemożliwa – jednak istotne staje się ujawnienie momentu, w którym architektura przestaje ułatwiać decyzję, a zaczyna systematycznie sterować zachowaniem.

W tym punkcie sztuczna inteligencja spotyka się z ekonomią uwagi. Jeśli podstawowym modelem biznesowym jest maksymalizacja czasu, zaangażowania lub konwersji, system uczący się będzie szukał coraz skuteczniejszych sposobów utrzymania użytkownika. Nie trzeba przy tym zakładać złej intencji projektanta; wystarczy sama funkcja celu. Jeśli oburzenie i konflikt silniej zwiększają zaangażowanie niż spokojna analiza, optymalizacja może premiować treści polaryzujące bez wydawania komukolwiek bezpośredniego polecenia: „polaryzuj społeczeństwo”. Jest to klasyczny przykład emergentnej niezgodności między miernikiem ekonomicznym a dobrem społecznym.

Problem AI jest zatem również problemem modelu biznesowego. To, czy system służy użytkownikowi, reklamodawcy, właścicielowi infrastruktury, pracodawcy czy państwu, determinuje jego funkcję celu. „Asystent użytkownika” finansowany przez podmiot trzeci może cierpieć na

konflikt lojalności, podobnie jak pośrednik w każdej innej dziedzinie. Prawdziwa zgodność z interesem człowieka wymaga więc nie tylko odpowiednich danych treningowych, lecz właściwej architektury bodźców ekonomicznych i odpowiedzialności prawnej. Z tej perspektywy szczególnie ciekawa jest możliwość rozwoju AI działającej jako fiducjarny agent jednostki – narzędzia zobowiązanego do działania w najlepiej pojętym interesie użytkownika, a nie do maksymalizacji korzyści podmiotu kontrolującego platformę. Taka idea przenosi problem alignmentu z poziomu abstrakcyjnego „dopasowania do ludzkości" na znany język prawa: obowiązek lojalności, ujawnianie konfliktów interesów, poufność i odpowiedzialność. Nie rozwiązuje to wszystkich problemów, ale pokazuje, że governance AI może czerpać nie tylko z informatyki, lecz z wielowiekowego dorobku regulowania relacji agencyjnych. Nowość tych wyzwań bywa bowiem przeceniana.

AI radykalizuje pytania znane prawu spółek, medycynie, administracji, finansom i etyce zawodowej: kto reprezentuje czyj interes, kto odpowiada za decyzję, jakie informacje musi ujawnić profesjonalista, kiedy konflikt interesów jest niedopuszczalny, jak zapewnić możliwość odwołania i kto ponosi koszt szkody.

Nowa definicja kompetencji i warunki emancypacyjnej automatyzacji

Technologiczna nowość dotyczy skali i szybkości, lecz normatywny rdzeń problemu pozostaje często niezmienny. Największa transformacja może jednak objąć samo rozumienie kompetencji. W świecie przed AI kompetencja oznaczała przede wszystkim zdolność wykonania zadania; w erze AI będzie oznaczać także umiejętność właściwego delegowania, oceniania i korygowania pracy systemu. Człowiek kompetentny nie musi generować wariantów szybciej niż model, ale powinien wiedzieć, kiedy mu nie ufać, jakie źródło zweryfikować, jak sformułować ograniczenia i jak rozpoznać odpowiedź pozornie elegancką, lecz błędną. Może to prowadzić do nowego rodzaju nierówności: różnic w zdolności korzystania z inteligencji zewnętrznej. Osoba z wysoką wiedzą dziedzinową wykorzysta AI jako narzędzie amplifikujące kompetencję; osoba pozbawiona tej wiedzy może nie odróżnić poprawnego wyniku od przekonującej konfabulacji. Ta sama technologia może zatem demokratyzować dostęp do pewnych czynności, jednocześnie zwiększając premię za zdolność ich krytycznej oceny. Jest to istotna korekta wobec naiwnego przekonania, że powszechna AI automatycznie wyrówna kapitał poznawczy. Dostęp do narzędzia nie jest tożsamy ze zdolnością jego efektywnego wykorzystania. Tak jak powszechny dostęp do bibliotek nie zniósł różnic

edukacyjnych, tak dostęp do modeli nie musi znosić różnic w wiedzy, rozumowaniu i sprawstwie – może jedynie zmienić ich strukturę.

AI przywraca nas do capability approach. Kluczowe pytanie nie brzmi, czy system wykonał za człowieka więcej, lecz czy dzięki niemu człowiek zyskał szerszy zbiór realnych możliwości. Dobrze zaprojektowana AI pozwala osobie z niepełnosprawnością komunikować się, małej organizacji analizować dokumenty dotąd zarezerwowane dla wielkich kancelarii, uczniowi otrzymać indywidualne wyjaśnienie, naukowcowi przeszukać obszerniejszy korpus literatury, a przedsiębiorcy zautomatyzować rutynę. W tych przypadkach technologia rozszerza podmiotowość. Ta sama technologia może jednak redukować capability, gdy człowiek przestaje rozumieć proces, traci umiejętność działania bez infrastruktury, nie potrafi zakwestionować wyniku lub zostaje podporządkowany modelowi oceny kontrolowanemu przez inny podmiot. Odpowiedź na pytanie, czy AI „wzmacnia" człowieka, zależy więc od architektury relacji człowiek-system. Pozwala to sformułować bardziej wymagające kryterium human-centered AI.

System wzmacnia podmiotowość wtedy, gdy zwiększa zdolność działania, nie odbierając jednocześnie zdolności rozumienia celu, kontroli środków, sprzeciwu wobec wyniku oraz możliwości opuszczenia relacji z systemem. Jeśli wszystkie cztery warunki zostaną spełnione, automatyzacja może być emancypacyjna. Jeśli któryś z nich zaniknie, korzyść funkcjonalna zostanie okupiona zależnością. Szczególnie istotna jest ostatnia z tych zdolności – możliwość wyjścia. System, z którego nie można praktycznie zrezygnować, posiada władzę większą niż zwykłe narzędzie. Dlatego interoperacyjność, przenoszalność danych, konkurencja i dostęp do alternatyw nie są jedynie technicznymi szczegółami polityki cyfrowej. Są elementami architektury wolności.

Pluralizm instytucjonalny jako realna forma alignmentu

Monopol epistemiczny jest niebezpieczny niezależnie od tego, czy sprawuje go państwo, korporacja czy zdecentralizowana infrastruktura. W tym świetle demokratyczne zarządzanie AI nie powinno sprowadzać się do głosowania większości obywateli nad parametrami modelu. Demokracja oznacza raczej, że zasady określające dopuszczalne zastosowania technologii, prawa jednostki, odpowiedzialność oraz granice koncentracji władzy podlegają publicznemu uzasadnieniu i mogą być zmieniane.

Eksperci są niezbędni do ustalania faktów technicznych, lecz sama ekspertyza nie daje im wyłącznego prawa do definiowania wartości. Technik może wyjaśnić, co system potrafi, ale nie oznacza to, że posiada on mandat do rozstrzygnięcia, czy społeczeństwo powinno z tego rozwiązania korzystać. To jedno z kluczowych rozróżnień między technokracją a demokracją

epistemicznie dojrzałą. Demokracja nie musi odrzucać wiedzy specjalistycznej; musi natomiast rozdzielić autorytet epistemiczny od autorytetu politycznego. Epidemiolog posiada kompetencje do oceny transmisji choroby, informatyk do dbania o bezpieczeństwo systemu, a ekonomista do analizy bodźców – żaden z nich nie otrzymuje jednak wraz z wiedzą automatycznego mandatu do ważenia konkurujących praw i wartości. AI uwypukla ten problem w sposób szczególny, gdyż podmioty dysponujące największą wiedzą techniczną są często jednocześnie graczami rynkowymi rozwijającymi tę technologię.

Istniejący konflikt interesów nie sprawia, że ekspertyza staje się nieważna. Oznacza on jedynie, że publiczny system oceny nie może opierać się wyłącznie na samoopisie branży. Wymaga niezależnego audytu, nauki, organizacji społecznych, regulatorów, konkurencyjnych ośrodków wiedzy oraz procedur zapewniających dostęp do danych niezbędnych do oceny skutków. Dopiero w takim pluralistycznym ekosystemie „radykalna racjonalność" może oznaczać coś więcej niż tylko wezwanie do rozsądku. Racjonalność instytucjonalna polega na stworzeniu warunków, w których twierdzenie można sprawdzić, błąd ujawnić, interes nazwać, decyzję zakwestionować, a politykę zmienić w świetle nowych dowodów. Żaden pojedynczy człowiek, model czy regulator nie musi być nieomylny, o ile system społeczny jest sprawnie zorganizowany pod kątem korekcji błędu. Właśnie to może stanowić najbardziej realistyczną wersję alignmentu.

Być może nie istnieje możliwość jednorazowego „dopasowania AI do wartości ludzkości", gdyż wartości są pluralistyczne, konflikty trwałe, a przyszłe zastosowania nieprzewidywalne. Możemy jednak budować instytucje, w których systemy pozostają korygowalne, audytowalne, ograniczone prawem, konkurencyjne i podporządkowane procedurom odpowiedzialności. Alignment staje się wówczas nie pojedynczym problemem programistycznym, lecz ciągłym procesem polityczno-instytucjonalnego dopasowywania technologii do zmieniającej się, choć nie dowolnej sfery wartości. W takim ujęciu celem nie jest zamrożenie sztucznej inteligencji na poziomie czystego narzędzia.

Ochrona funkcji konstytutywnych jako warunek podmiotowości

Narzędzia zawsze zmieniały swoich użytkowników. Pismo przekształciło pamięć, druk strukturę autorytetu, samochód geografię miast, a telefon znaczenie odległości. AI również zmieni nasze praktyki poznawcze. Kluczowe jest jednak to, czy zmiana ta pozostanie refleksyjna – czy społeczeństwo zdoła obserwować, jakie kompetencje zyskuje i traci, komu przekazuje sprawstwo, gdzie powstają nowe asymetrie i które funkcje chce zachować jako konstytutywne dla osoby.

Prowadzi to do subtelniejszego pytania niż klasyczne: „czy AI zastąpi człowieka?”. Człowiek może zostać „zastąpiony” bez zniknięcia z miejsca pracy, jeśli stanie się jedynie nominalnym akceptantem decyzji. Może jednak przestać wykonywać część czynności i jednocześnie uzyskać większe sprawstwo, o ile technologia uwolni go od rutyny, pozwalając skupić się na celu, relacji i osądzie. Automatyzacja i podmiotowość nie są zatem prostymi przeciwieństwami; o ich relacji decyduje architektura delegacji.

Grayling formułuje ten problem w języku ostrzeżenia przed „jaźnią nominalną” i obrony „jaźni istotnej”. W dalszym wywodzie należy ten język precyzyjniej zrekonstruować, gdyż pojęcie nominalności obejmuje tu dwa odrębne zjawiska: symulację osoby przez system oraz stopniową nominalizację sprawstwa człowieka poprzez outsourcing poznania. Oba łączy jedna intuicja: zewnętrzne podobieństwo do działania podmiotowego nie gwarantuje rzeczywistego autorstwa.

W tym miejscu pojawia się pytanie: co właściwie powinniśmy chronić, mówiąc o ochronie człowieka przed nadmierną automatyzacją? Nie może chodzić o zachowanie każdej czynności wykonywanej ręcznie – taki konserwatyzm zatrzymałby historię techniki. Nie chodzi również o obronę biologicznej niezmienności. Bardziej przekonujące jest wskazanie funkcji konstytutywnych: zdolności rozumienia i kwestionowania, autorstwa własnych celów, odpowiedzialności, wzajemnego uznania oraz uczestnictwa w realnych relacjach. Jeśli te funkcje zostaną zachowane, AI może stać się jednym z najpotężniejszych narzędzi emancypacji poznawczej w historii. Jeśli zostaną wydrążone, technologia stworzy świat sprawnych procesów, w którym człowiek coraz częściej podpisuje decyzje, których nie podjął, powtarza sądy, których nie zweryfikował i realizuje cele, których sam nie ustanowił, pozostając formalnie wolnym w środowisku precyzyjnie zarządzającym jego prawdopodobnym zachowaniem.

Problem sztucznej inteligencji nie daje się zamknąć w technofilii czy technofobii. AI jest wzmacniaczem: potrafi multiplikować wiedzę i błędy, sprawstwo i zależność, pluralizm i centralizację, transparentność i manipulację. Jej kierunek nie wynika automatycznie z architektury modelu, lecz powstaje w układzie systemu, instytucji, prawa, rynku, kultury i ludzkich nawyków. Przyszłość AI będzie więc w znacznym stopniu przyszłością podziału sprawstwa. Pytanie nie brzmi, czy człowiek albo maszyna będą „rządzić”, gdyż rzeczywiste systemy będą hybrydowe.

Należy pytać, w których punktach procesu człowiek ustanawia cel, gdzie system wybiera środki, kto kontroluje dane, kto może zakwestionować wynik, kto odpowiada za szkodę i czy obywatel ma realną możliwość odmowy. W tych pozornie technicznych detalach ukrywa się konstytucja przyszłego społeczeństwa. Dlatego najcenniejszą cechą AI wcale nie musi być maksymalna autonomia, lecz dobrze zaprojektowana zależność od człowieka: system wystarczająco samodzielny, by zwiększać ludzkie możliwości, lecz dostatecznie przejrzysty i korygowalny, by nie stać się

arbitrem. Najbardziej rozwinięta technologia to nie ta, która usuwa człowieka z procesu, lecz ta, która rozumie, w których miejscach nie wolno go usunąć.

Jeśli tak, kolejnym problemem przestaje być samo bezpieczeństwo systemu, a staje się nim ontologia podmiotowości. Trzeba zapytać, czym różni się człowiek korzystający z AI od człowieka, którego poznanie zostało przez nią zastąpione; czym symulowana rozmowa różni się od relacji; czy kompetencja językowa wystarcza do przypisania podmiotowości oraz jaką rolę odgrywają świadomość, biografia, cielesność, podatność na zranienie i możliwość odpowiedzialności. Dopiero po tym rozróżnieniu będzie można ocenić, czy „jaźń istotna" i „jaźń nominalna" są adekwatnymi kategoriami do opisu nadchodzącego świata – i co właściwie znaczy zachować człowieka w systemie, który coraz lepiej symuluje jego głos.

> Czy w świecie, który perfekcyjnie symuluje ludzki głos i procesy myślowe, będziemy jeszcze w stanie odróżnić kompetencję od imitacji, a realne sprawstwo od rytualnego zatwierdzania cudzych decyzji? Być może największym wyzwaniem nie jest stworzenie bezpiecznej maszyny, lecz zdefiniowanie na nowo, co w nas pozostaje niezbywalne i niepodlegające outsourcingowi. Ostatecznie o naszej podmiotowości nie zdecyduje to, co AI potrafi za nas zrobić, ale to, w których miejscach kategorycznie odmówimy jej zastąpienia nas samych.

Inspiracje

[1] "The Challenge of the Future" A C Grayling

Anthony Clifford Grayling (ur. 1949) to brytyjski filozof, pisarz i intelektualista. Jest założycielem i byłym magistrem New College of the Humanities (obecnie Northeastern University London) oraz członkiem nadzwyczajnym St Anne's College w Oksfordzie. Wcześniej był profesorem filozofii na Birkbeck University of London. Dorobek naukowy Graylinga obejmuje epistemologię, logikę filozoficzną, etykę humanistyczną i historię idei. Jako wybitny obrońca sekularyzmu, praw człowieka i zasad demokratycznych, napisał i zredagował ponad trzydzieści książek oraz publikował liczne artykuły w gazetach, czasopismach i mediach. Jego badania naukowe łączą filozofię techniczną z dyskursem publicznym, zgłębiając fundamentalne dylematy moralne i tradycje filozoficzne w kontekście współczesnych problemów globalnych.

Anthony Clifford Grayling (born 1949) is a British philosopher, author, and public intellectual. He is the founder and former Master of the New College of the Humanities (now Northeastern University London) and a supernumerary fellow of St Anne's College, Oxford. Previously, he served as Professor of Philosophy at Birkbeck, University of London. Grayling's academic work encompasses epistemology, philosophical logic, humanist ethics, and the history of ideas. A prominent public defender of secularism, human rights, and democratic principles, he has written and edited more than thirty books and contributed widely to newspapers, journals, and broadcast media. His scholarship bridges technical philosophy and public discourse, exploring fundamental moral dilemmas and philosophical traditions to address contemporary global issues.

Northeastern University London

Literatura uzupełniająca

Książki

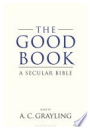
Książki autora i dzieła przywołane w tekście:



[1] "Grace"

A C Grayling

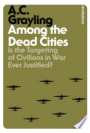
Oberon Books



[2] "The Good Book"

A C Grayling

A&C Black | ISBN: 9780747599609



[3] "Among the Dead Cities"

A C Grayling

A&C Black | ISBN: 9781472534057



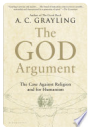
[4] "The Meaning of Things"

A C Grayling

Weidenfeld & Nicolson | ISBN: 9781780221168

[5] ""Bomber County": A riveting story of the air war"

A C Grayling



[6] "The God Argument"

A C Grayling

Bloomsbury+ORM | ISBN: 9781620401910

Książki autorów przywoływanych i literatura pokrewna:



[7] "Human Enhancement"

Nick Bostrom

OUP Oxford | ISBN: 9780191559600



[8] "Ergonomics and Human Factors"

Purnima Embranthiri

2024 | Educohack Press | ISBN: 9789361524905

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "IEEE 3rd Conference on Norbert Wiener in the 21st Century: Being Human in a Global Village"

N. Wiener, G. Adamson

2020 | IEEE technology & society magazine | DOI: 10.1109/mts.2020.3002051

Źródło

[2] **"Cognitive Enhancement: Methods, Ethics, Regulatory Challenges"**

Nick Bostrom, Anders Sandberg

2009 | Science and Engineering Ethics | DOI: 10.1007/s11948-009-9142-5

Źródło

[3] **"Future Progress in Artificial Intelligence: A Survey of Expert Opinion"**

Vincent C. Müller, Nick Bostrom

2016 | Synthese Library/Synthese library | DOI: 10.1007/978-3-319-26485-1_33

Źródło

[4] **"Existential risks: analyzing human extinction scenarios and related hazards"**

Nick Bostrom

2002 | Oxford University Research Archive (ORA) (University of Oxford)

Źródło

[5] **"Human Genetic Enhancements: A Transhumanist Perspective"**

Nick Bostrom

2003 | The Journal of Value Inquiry | DOI: 10.1023/b:inqu.0000019037.67783.d5

Źródło

[6] **"The Order of Things: An Archaeology of the Human Sciences."**

F. C. T. Moore, Michel Foucault

1971 | Man | DOI: 10.2307/2799252

Źródło

[7] **"The future as rupture: Michel Foucault"**

Michel Foucault

2011 | Time and Philosophy | DOI: 10.1017/upo9781844652778.013

Źródło

[8] **"Sustainable Human Development: Concepts and Priorities"**

Sudhir Anand and Amartya Sen

1994 | RePEc: Research Papers in Economics

Źródło

[9] **"Gender Inequality in Human Development: Theories and Measurement"**

Sudhir Anand and Amartya Sen

1995 | RePEc: Research Papers in Economics

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: 9cda81f5-b8a1-475f-8193-5fa71e62c574