Poza AGI: Od inteligencji maszynowej do hybrydowej cywilizacji poznawczej w świetle koncepcji The AI Instinct Rany Gujral

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł kwestionuje liniowy model rozwoju sztucznej inteligencji, w którym AGI jest jedynie stopniem do ASI. Autor proponuje przejście od analizy samych kompetencji do koncepcji Artificial General Experience (AGE), która wprowadza wymiar czasu, kontekstu i pamięci. Tekst podkreśla fundamentalne różnice między biologiczną kognicją człowieka – zakorzenioną w ewolucji i homeostazie – a matematyczno-obliczeniową naturą AI. Zamiast szukać homologii mechanizmów, należy skupić się na funkcjonalnej komplementarności obu systemów. Rozwój AI jest przedstawiony jako wielowymiarowa ewolucja, prowadząca nie do zastąpienia człowieka, lecz do powstania hybrydowej cywilizacji poznawczej, w której kluczowe staje się pytanie o interoperacyjność infrastruktury myślowej i zachowanie suwerenności poznawczej.

The article questions the linear model of artificial intelligence development, in which AGI is merely a stepping stone to ASI. The author proposes a shift from analyzing competencies alone to the concept of Artificial General Experience (AGE), which introduces the dimensions of time, context, and memory. The text emphasizes the fundamental differences between human biological cognition—rooted in evolution and homeostasis—and the mathematical-computational nature of AI. Instead of seeking homology between mechanisms, focus should be placed on the functional complementarity of both systems. The development of AI is presented as a multidimensional evolution leading not to the replacement of humans, but to the emergence of a hybrid cognitive civilization, where the key question becomes the interoperability of thought infrastructure and the preservation of cognitive sovereignty.

Artykuł rzuca wyzwanie dominującemu przekonaniu, że rozwój sztucznej inteligencji jest liniowym marszem ku AGI i ASI. Autor stawia tezę, że przyszłość AI to nie wzrost pojedynczej kompetencji, lecz ewolucja w stronę wielowymiarowej cywilizacji hybrydowej, w której kluczowe staje się rozróżnienie między funkcjonalną sprawnością maszyn a fenomenalnym doświadczeniem człowieka. Tekst analizuje ryzyka związane z „dekapitalizacją kompetencji" – sytuacją, w której wzrost bieżącej produktywności dzięki AI odbywa się kosztem utraty ludzkiej suwerenności poznawczej i zdolności do niezależnego myślenia. Zamiast wizji fuzji umysłów, autor proponuje model „rozszerzonej biografii", w której człowiek pozostaje centrum doświadczenia, ale korzysta z zewnętrznej infrastruktury kognitywnej. Głównym postulatem jest ochrona wolności epistemicznej poprzez pluralizm systemów i ramy konstytucyjne, które zapobiegną arbitralnej dominacji technologicznej i „kalcyfikacji" ludzkiego ducha w klatce algorytmicznej optymalizacji.

SŁOWA KLUCZOWE

hybrydowa cywilizacja poznawcza | Artificial General Experience (AGE) | suwerenność poznawcza | dekapitalizacja kompetencji | kognicja hybrydowa | model Centaura | cognitive offloading | prawo wymaganej różnorodności Ashby'ego | rozszerzona biografia | nieciągłość predykcyjna | metakognicja społeczna | biurokracja kognitywnie aktywna | interoperacyjność infrastruktury poznawczej | alignment systemowy

Rozwój AI to wielowymiarowa ewolucja, a nie liniowy wzrost kompetencji

Co poza AGI? Od inteligencji maszynowej do hybrydowej cywilizacji poznawczej. Pytanie o to, co znajduje się poza AGI, wydaje się początkowo pytaniem o kolejny stopień technologicznej drabiny: AGI miałaby zostać zastąpiona przez ASI, a następnie przez jeszcze doskonalszą architekturę, tak jakby historia sztucznej inteligencji przypominała liniowy marsz przez kolejne poziomy kompetencji.

W rzeczywistości sprawa jest znacznie bardziej złożona. AGI stanowi przede wszystkim próbę określenia szerokości i poziomu kompetencji systemu, a nie kompletną teorię kognicji, doświadczenia, podmiotowości czy odpowiedzialności, ani analizę społecznych konsekwencji automatyzacji. Nawet współczesne ramy operacjonalizujące AGI oddzielają generalność kompetencji od ich poziomu, traktując autonomię jako wymiar wdrożenia, którego nie należy utożsamiać z samą inteligencją.

Tym samym przekroczenie kolejnego benchmarku nie odpowiada na pytania o ciągłość pamięci, uczenie się na błędach, społeczne umocowanie decyzji czy zdolność do rozpoznawania własnej niewiedzy, a tym bardziej nie rozstrzyga, kto powinien ponosić koszt pomyłek systemu. Właśnie przeciwko redukcji przyszłości AI do osi kompetencji formułowana jest koncepcja Artificial General Experience (AGE). AGE ma stanowić „brakującą oś", na której znaczenie uzyskują czas, kontekst, pamięć oraz konsekwencje zmieniające przyszłe zachowanie systemu. Jest to intelektualnie płodna propozycja, o ile zachowamy rozróżnienie wypracowane w poprzednich częściach, gdyż AGE nie jest obecnie utrwalonym terminem naukowym, a samo „doświadczenie" posiada co najmniej dwa odmienne znaczenia.

Z jednej strony system może posiadać historię, długotrwałą pamięć i zdolność do zmiany polityki po błędzie, rekonstruując model świata w oparciu o konsekwencje wcześniejszych działań. Z drugiej strony doświadczenie oznacza fenomenalne przeżywanie – istnienie stanu, w którym „coś znaczy", jak to jest być danym systemem. Pierwszy problem ma charakter głównie inżynierski i badawczy; drugi pozostaje związany z nierozstrzygniętym problemem świadomości. Nie wolno przejść od funkcjonalnej historyczności do qualiów za pomocą samej zmiany słownika. To rozróżnienie pozwala zdemitologizować pytanie o „następcę AGI". Jeśli system stanie się bardziej generalny, niezawodny i autonomiczny, może uzyskać szerokie kompetencje bez dowodu posiadania fenomenalnego życia. Jeśli otrzyma trwałą pamięć i będzie adaptował się do skutków własnych działań, zyska historyczność funkcjonalną bez konieczności cierpienia. Nawet jeśli zostanie wyposażony w robotyczne ciało i uzyska sensorimotoryczną pętlę ze środowiskiem, ucieleśnienie techniczne nadal nie będzie automatycznym odpowiednikiem biologicznej homeostazy. Stworzenie self-modelu nie oznacza jeszcze powstania fenomenalnego „ja".

Rozwój AI należy więc przedstawiać jako przestrzeń wielu częściowo niezależnych wymiarów, a nie jako jedną skalę prowadzącą od kalkulatora do człowieka, a następnie do boga-maszyny. Takie wielowymiarowe podejście odpowiada temu, czego nauczyły nas nauki o ludzkiej kognicji. Inteligencji człowieka nie da się opisać jedną liczbą oderwaną od środowiska, historii i biologii. Decyzje powstają w warunkach ograniczonej racjonalności Herberta Simona, selektywnej uwagi, ograniczeń pamięci roboczej, heurystyk, ucieleśnionego wartościowania, emocji, motywacji,

interocepcji, norm społecznych i języka. Mózg nie jest uniwersalnym procesorem oczekującym na problem w czystej formie, lecz żywym układem regulacyjnym, który przez miliony lat ewolucji rozwijał zdolność działania pod presją czasu, niedoboru informacji i kosztów biologicznych. Jego racjonalność jest zatem ekologiczna i sytuacyjna; zachowanie człowieka może być suboptymalne wobec abstrakcyjnego benchmarku, a jednocześnie w pełni adaptacyjne wobec realnych kosztów środowiska.

Sztuczna kognicja posiada inną genealogię. Modele uczenia maszynowego nie są niedojrzałymi ludźmi, lecz systemami matematyczno-obliczeniowymi budowanymi według innych zasad, na innych substratach i z innymi ograniczeniami pamięci czy energii. Ich podobieństwo do zachowań człowieka może być funkcjonalne bez homologii mechanizmów. To dlatego odrzucono zbyt łatwą analogię, według której obecne modele byłyby „czystym Systemem 2", a człowiek miałby dostarczać im Systemu 1, emocji albo intuicji. Badania nad LLM pokazują zachowania przypominające zarówno szybkie heurystyki, jak i procedury deliberacyjne, lecz dual-process theory nie może być bezpośrednio przenoszona z psychologii człowieka na architekturę sztuczną jako opis mechanistycznie tożsamy.

Najbardziej produktywna analiza zaczyna się wtedy, gdy zamiast pytać, jak wiele człowieka znajduje się w maszynie, pytamy, jakie funkcje realizuje każdy typ systemu i gdzie ich odmienności mogą się wzajemnie uzupełniać. Z tej perspektywy „poza AGI" pojawia się pierwsza zasadnicza zmiana skali: przejście od inteligencji izolowanego systemu do inteligencji konfiguracji człowiek-AI, opisanej w książce formułą Human + Tools + Rules oraz typologią Centaur, Cyborg i Hive.

Wydajność hybrydowa zależy od projektu sprzężenia, a nie sumy kompetencji

Wartość tej perspektywy nie tkwi w futurystycznej efektowności, lecz w przesunięciu jednostki analizy. W realnym środowisku społecznym wynik nie jest produktem modelu pracującego w próżni; powstaje on w układzie obejmującym dane, człowieka, interfejs, organizację, prawo, kryteria sukcesu, procedury kontroli oraz historię wcześniejszych decyzji. To właśnie ten układ powinien być oceniany pod kątem jakości decyzji, odporności na błąd, dystrybucji sprawczości i zdolności do uczenia się. Dane empiryczne stanowią tutaj istotne antidotum na naiwny technologiczny optymizm.

Metaanaliza 106 badań eksperymentalnych obejmujących 370 efektów wykazała, że zespoły człowiek-AI osiągały przeciętnie wyniki lepsze niż sami ludzie, lecz jednocześnie gorsze niż

najlepszy z dwóch komponentów działających osobno. Straty te były szczególnie widoczne w zadaniach decyzyjnych, podczas gdy potencjał komplementarności okazywał się większy w zadaniach twórczych. Wynik ten podważa prostą arytmetykę postępu: nie wystarczy połączyć inteligentnego człowieka z kompetentnym modelem, aby automatycznie uzyskać „więcej inteligencji". Kluczowy jest projekt sprzężenia – podział ról, kolejność dostępu do informacji, kalibracja zaufania, epistemiczna niezależność, możliwość kwestionowania odpowiedzi oraz zgodność mechanizmu delegowania z charakterem zadania. W tym punkcie cybernetyka dostarcza języka niezwykle użytecznego dla całej teorii Hybrydowej Ludzkości.

Norbert Wiener zwrócił uwagę na znaczenie sprzężeń zwrotnych, a W. Ross Ashby sformułował prawo wymaganej różnorodności, zgodnie z którym zdolność regulatora do kontrolowania złożonego środowiska zależy od jego własnego repertuaru możliwych odpowiedzi. W kontekście AI oznacza to, że samo zwiększanie „inteligencji" jednego modułu jest niewystarczające. System musi posiadać odpowiednią różnorodność reakcji, procedur eskalacji, niezależnych źródeł oraz sposobów korygowania błędów, aby sprostać złożoności napotykanych sytuacji. Monolityczna AI może być niezwykle wydajna w przestrzeni przewidywanej przez konstruktorów, a jednocześnie krucha wobec problemów wymagających alternatywnej reprezentacji.

Z teorią Ashby'ego koresponduje Herbert Simon z ideą systemów hierarchicznych i near decomposability. Złożone układy są często odporne właśnie dlatego, że nie wszystkie ich elementy są ze sobą równie silnie połączone. Moduły zachowują częściową autonomię, a interakcje między nimi są wystarczające do koordynacji, lecz nie na tyle intensywne, by każda lokalna perturbacja natychmiast destabilizowała całość. Dla przyszłej kognicji hybrydowej oznacza to, że maksimum integracji nie musi być optimum. Fused Mind, rozumiany dosłownie jako eliminacja granic między komponentami, może okazać się mniej odporny niż system zachowujący modularność, redundancję i epistemiczną niezależność. Czasem granica między człowiekiem a AI powinna być intencjonalnie utrzymywana właśnie po to, by możliwe było wzajemne wykrywanie błędów. Model Centaura nie jest zatem niedoskonałym stadium przed pełnym Cyborgiem; w niektórych domenach wysokiej stawki może być architekturą dojrzalszą.

Lekarz, prawnik, inżynier czy analityk bezpieczeństwa powinien czasami sformułować własną hipotezę przed zapoznaniem się z sugestią systemu, gdyż niezależność poznawcza ma wartość diagnostyczną. Jeśli AI pierwsza definiuje problem, człowiek staje się podatny na kotwiczenie i automation bias. Jeżeli natomiast oba komponenty rozpoczynają pracę od częściowo odrębnych reprezentacji, ich niezgoda może ujawnić przypadek wymagający szczególnej analizy. Komplementarność nie powstaje więc przez zacieranie różnic, lecz przez ich inteligentne wykorzystanie. W innym rodzaju zadań wartościowa może być głębsza integracja.

Pisanie, projektowanie, programowanie, planowanie i eksploracja hipotez mogą korzystać z bardzo krótkich pętli człowiek-AI, w których trudno przypisać poszczególne elementy finalnego rezultatu jednemu autorowi procesu poznawczego. Tak powstaje Cyborg w znaczeniu funkcjonalnym: nie jako nowa fenomenalna osoba, lecz proces, którego wydajność i kształt zależą od iteracyjnej współadaptacji. Człowiek koryguje AI, AI modyfikuje kolejne pomysły człowieka, a obie strony uczą się skuteczniejszych sposobów komunikacji. Z czasem wykształca się kompetencja relacyjna, której nie da się wyjaśnić jedynie statycznymi parametrami modelu lub wiedzą użytkownika. Właśnie tutaj cognitive offloading staje się zjawiskiem ambiwalentnym.

Zewnętrzne narzędzia od zawsze pozwalały ludziom przesuwać operacje poznawcze do środowiska, zwiększając zdolność rozwiązywania problemów. Najnowsze przeglądy potwierdzają, że offloading może znacząco podnieść efektywność, jednak jego konsekwencje zależą od metakognicji użytkownika i sposobu, w jaki technologia uczestniczy w procesie uczenia. Jeśli AI zastępuje pamięć roboczą lub wykonuje rutynowe przeszukiwanie, może uwolnić zasoby do bardziej zaawansowanej refleksji. Jeżeli jednak przejmuje również generowanie hipotez, krytykę, ocenę i końcowy wybór, użytkownik może utracić praktykę niezbędną do niezależnego wykonywania tych funkcji. Problem nie leży zatem w outsourcingu jako takim, lecz w tym, czy zwiększa on, czy zmniejsza przyszłą zdolność podmiotu do sterowania własnym outsourcingiem.

Dekapitalizacja kompetencji a utrata suwerenności poznawczej

To rozróżnienie wprowadza nas w obszar kapitału poznawczego. Kompetencja nie jest bowiem jedynie bieżącą zdolnością do uzyskania wyniku, lecz przede wszystkim trwałą umiejętnością podmiotu do rozpoznania problemu, krytycznej oceny cudzej odpowiedzi i sprawnego działania w sytuacji awarii infrastruktury. Jeśli duet człowieka i AI osiąga znakomite rezultaty, ale człowiek z każdym rokiem traci zdolność do weryfikacji działań systemu, produktywność bieżąca rośnie kosztem przyszłej odporności.

Ekonomia określiłaby to jako konsumowanie kapitału w celu zwiększenia aktualnego przepływu. W kontekście poznawczym mówimy o dekapitalizacji kompetencji. Stąd jedno z kluczowych kryteriów dla systemów human-centered AI: narzędzie powinno być oceniane nie tylko przez pryzmat wydajności człowieka podczas jego używania, ale również przez retained human capability po długim okresie współpracy.

Ta sama zasada dotyczy edukacji, medycyny, prawa czy inżynierii. Uczeń oddający idealną pracę dzięki AI niekoniecznie nabył kompetencję. Lekarz stawiający trafną diagnozę z pomocą modelu może nie wykształcić zdolności rozpoznawania przypadków, w których ten model zawodzi. Programista wydajny dzięki agentowi może mieć mniejszą zdolność rekonstrukcji całego systemu niż poprzednie pokolenia specjalistów. W społeczeństwie hybrydowym musimy zatem odróżnić posiadanie zdolności od posiadania dostępu do zdolności – rozróżnienie to będzie miało fundamentalne znaczenie ekonomiczne.

Część ludzkiej produktywności może przestać być kapitałem ucieleśnionym w pracowniku, a stać się dzierżawioną funkcją dostępną jedynie poprzez system należący do pracodawcy lub platformy. Człowiek może być w pracy znacznie bardziej wydajny, a jednocześnie posiadać mniejszą przenośną siłę negocjacyjną. Opuszczając organizację, traci nie tylko oprogramowanie, ale i pamięć, agentów, personalizację oraz workflow współtworzące jego realną kompetencję. Technologia może zatem równocześnie zwiększać produktywność i rekonfigurować relacje między kapitałem, pracą a własnością. Właśnie dlatego analizy „poza AGI" nie da się prowadzić w oderwaniu od ekonomii politycznej. Im głębiej sztuczna infrastruktura przenika proces poznawczy, tym bardziej palące staje się pytanie o to, kto ją posiada i kto ustala jej reguły. Rozszerzony umysł Andy'ego Clarka i Davida Chalmersa nabiera w epoce chmury wymiaru, którego klasyczny eksperyment z notesem Ottona nie musiał obejmować.

Papierowy notes pozostawał pod kontrolą użytkownika. Cyfrowa pamięć bywa przechowywana na cudzych serwerach; podlega licencji, zmianom regulaminu, zdalnym aktualizacjom i ekonomii abonamentowej. Element funkcjonalnie tożsamy z ludzką pamięcią może zatem pozostawać własnością innego podmiotu. Sytuacja ta wymaga rozszerzenia pojęcia prywatności o suwerenność poznawczą. Nie jest ona postulatem powrotu do człowieka samowystarczalnego, gdyż taki nigdy nie istniał – Homo sapiens od zawsze funkcjonował w zależności od języka, innych ludzi, narzędzi i instytucji.

Suwerenność poznawcza jako brak arbitralnej dominacji infrastruktury

Suwerenność należy rozumieć raczej jako zdolność zarządzania zależnościami: możliwość wyboru systemu, przeniesienia pamięci, kontroli pełnomocnictw, ograniczenia zakresu personalizacji oraz uzyskania alternatywy i opuszczenia infrastruktury bez nieproporcjonalnej utraty własnej funkcjonalności. W języku republikańskiej teorii wolności problemem nie jest istnienie relacji

zależności samych w sobie, lecz brak arbitralnej dominacji. To właśnie republikańskie rozumienie wolności dostarcza trafnej korekty dla konsumenckiej wizji AI.

Użytkownik może przez lata nie doświadczać żadnej konkretnej ingerencji dostawcy i mimo to pozostawać w relacji dominacji, jeśli podmiot ten posiada jednostronną możliwość odebrania dostępu do krytycznej pamięci lub radykalnej zmiany działania systemu, a użytkownik nie dysponuje realną alternatywą. Wolność wymaga zatem nie tylko poprawnego aktualnego zachowania infrastruktury, lecz instytucjonalnego ograniczenia jej potencjalnej arbitralności. Przenośność, interoperacyjność, audyt, prawo do odwołania i strategie wyjścia stają się w tej perspektywie instrumentami ochrony sprawczości.

Ten sam problem rośnie wykładniczo przy przejściu z poziomu jednostki na poziom organizacji i państwa. Instytucja może dysponować ogromną ilością danych, a mimo to utracić część własnej zdolności działania, jeżeli kluczowe mechanizmy syntetyzowania wiedzy i koordynacji zależą od jednego zewnętrznego dostawcy. Państwo wykorzystujące AI do planowania, analizy ryzyka i administracji może zwiększyć swoją sprawność, lecz jednocześnie pogłębić zależność strategiczną od infrastruktury, nad którą nie sprawuje pełnej kontroli. „Suwerenność technologiczna” i „suwerenność poznawcza” okazują się tutaj pojęciami analogicznymi: nie wymagają autarkii, lecz realnej zdolności zmiany konfiguracji bez utraty ciągłości podstawowych funkcji.

Z ekonomii politycznej wynika kolejny wniosek: aby doszło do ogromnej asymetrii władzy, po AGI nie musi powstać jedna superinteligencja. Wystarczy, że określone organizacje jako pierwsze nauczą się budować sprawne hybrydowe organizmy decyzyjne. Przedsiębiorstwo wyposażone w rozbudowaną pamięć, sieć agentów, predykcję, szybki dostęp do danych oraz ludzi zdolnych kierować taką infrastrukturą może uzyskać przewagę wykraczającą poza klasyczną automatyzację. Państwo posiadające podobny system może z kolei radykalnie zwiększyć swoją zdolność administracyjną.

Nie będzie do tego konieczne stworzenie świadomego superkomputera. Wystarczy powstanie organizacji o znacznie większej zdolności zbiorowego poznania. To właśnie może być pierwsza realna postać superinteligencji: nie właściwość pojedynczego modelu, lecz właściwość instytucji. Nie jest to czysta spekulacja, gdyż już dziś organizacje funkcjonują jako systemy poznawcze rozproszone pomiędzy ludzi, procedury, dokumentację i aparaturę. AI zwiększa jedynie aktywność i przepustowość niebiologicznej części tego układu.

Ryzyko poznawczego konserwatyzmu w automatyzacji systemów

Archiwum przestaje być bierną pamięcią, skoro agent może przeszukiwać je przy każdej decyzji. Procedura przestaje być wyłącznie tekstem, gdyż system jest w stanie monitorować jej stosowanie. Rozproszone doświadczenie ekspertów można szybciej routować do konkretnego problemu, dzięki czemu organizacyjna kognicja staje się bardziej ciągła, aktywna i zdolna do samomonitorowania. Max Weber widział w biurokracji formę racjonalizacji opartą na specjalizacji, dokumentacji i formalnych regułach; agentowa AI może wprowadzić kolejną fazę tej logiki: biurokrację kognitywnie aktywną. Taka struktura nie tylko przechowuje regułę, ale interpretuje przypadek, porównuje go z tysiącami poprzednich, wykrywa odstępstwa, modeluje konsekwencje i uruchamia procedurę eskalacji. Z jednej strony zwiększa to zdolność organizacji do konsekwentnego działania, z drugiej zaś grozi „kalcyfikacją" norm – sytuacją, w której historyczny wzorzec jest automatycznie powielany tam, gdzie wymagana byłaby reinterpretacja. Tutaj ponownie pojawia się znaczenie ludzkiego i instytucjonalnego habit of refusal – zdolności stwierdzenia, że przypadek wymaga wyjścia poza to, co system zna najlepiej. Gujral właśnie ochronę takiego nawyku odmowy wiąże z obroną ludzkiej sprawczości.

Prawo stanowi wyjątkowo dobre laboratorium tego problemu. System uczony na wcześniejszych orzeczeniach może niezwykle skutecznie rozpoznawać wzorce rozstrzygnięć, jednak prawo nie jest jedynie statystycznym opisem własnej przeszłości. Rozwija się ono poprzez zmianę wykładni, konflikty wartości, nowelizacje, reinterpretację zasad konstytucyjnych i odpowiedź na nowe warunki społeczne. Algorytm optymalizowany pod kątem historycznej zgodności może zatem wytwarzać konserwatyzm poznawczy znacznie silniejszy od zamierzonego. Predykcja precedensu i tworzenie prawa są epistemicznie różnymi czynnościami: pierwsza może być w dużym stopniu automatyzowana, druga zaś wymaga normatywnej legitymizacji oraz zdolności uzasadnienia odejścia od przeszłości.

Podobny problem pojawia się w nauce. AI może zwiększyć sprawność przeszukiwania literatury, tworzenia hipotez i znajdowania wzorców w danych, ale nauka nie jest wyłącznie maszyną predykcyjną. Jej sukces zależy również od projektowania testów, krytyki założeń, rozróżniania korelacji od przyczynowości oraz utrzymywania teorii alternatywnych. Najbardziej wartościowy „Explanation Engine" nie powinien więc być systemem przedstawiającym jedną, perfekcyjnie płynną teorię rzeczywistości, lecz infrastrukturą generowania, porównywania, falsyfikowania i aktualizowania wielu modeli. Źródłowe pojęcie Silnika Wyjaśnień można w ten sposób uczynić naukowo defensywnym, nie przypisując AI uprzywilejowanego „bezpośredniego wglądu w naturę".

Filozofia nauki pomaga przy tym odróżnić predykcję, wyjaśnienie i rozumienie. Dobry model predykcyjny może przewidywać wynik bez reprezentowania mechanizmu przyczynowego w formie zrozumiałej dla człowieka. Wyjaśnienie wskazuje natomiast struktury i zależności pozwalające odpowiedzieć nie tylko na pytanie „co się wydarzy?", ale również „dlaczego?", „co zmieniłoby wynik?" i „w jakich warunkach model przestanie działać?". Rozumienie może obejmować jeszcze szerszą zdolność osadzenia mechanizmu w sieci innych pojęć i zastosowania go w sytuacjach nowych. W przyszłych hybrydach różne elementy mogą odpowiadać za poszczególne etapy tego procesu, a granica między „ludzkim rozumieniem" a „maszynową predykcją" prawdopodobnie stanie się mniej stabilna niż dzisiaj.

W tym miejscu pojawia się teoria informacji. Shannonowska informacja mierzy redukcję niepewności w kanale komunikacyjnym i nie jest sama w sobie teorią znaczenia. System może przesyłać i kompresować ogromną ilość informacji syntaktycznej bez rozwiązania filozoficznego problemu semantyki. To rozróżnienie jest kluczowe w epoce LLM, ponieważ rosnąca zdolność do modelowania statystycznej struktury języka może produkować bardzo wysoką kompetencję semantyczno-pragmatyczną bez konieczności przyjmowania fenomenalnej subiektywności. Przyszłość inteligencji wymaga więc odróżnienia ilości informacji, organizacji informacji, funkcjonalnego znaczenia i przeżywanego znaczenia.

Funkcjonalny koszt a fenomenalne doświadczenie

AGE próbuje przywrócić do tej struktury czas i konsekwencję. System nie powinien jedynie przechowywać rekordów w bazie danych, lecz stawać się inny pod wpływem tego, co wydarzyło się wcześniej. Ta intuicja znajduje silne analogie w teorii uczenia, plastyczności, pamięci organizacyjnej oraz biologii adaptacji. Organizm nie posiada doświadczenia dlatego, że zachował dane o przeszłości; to przeszłość zmienia jego reakcje, priorytety i sposób interpretacji nowych bodźców. Funkcjonalne AGE można zatem analizować jako zależność przyszłej polityki systemu od historii jego interakcji, bez konieczności przypisywania temu procesowi bólu czy fenomenalności.

W tym miejscu należy skorygować jedną z najmocniejszych tez książki, według której prawdziwa inteligencja wymagałaby "wewnętrznego kosztu", a błąd musiałby w pewnym sensie boleć, by prowadzić do autentycznej teleologii. W biologii koszt ma rzeczywiście znaczenie konstytutywne: organizm działa pod presją homeostazy, energii, uszkodzeń i śmierci. Nie wynika z tego jednak logiczna konieczność fenomenalnego cierpienia jako warunku adaptacji. System sztuczny może internalizować koszt funkcjonalnie poprzez zmianę wartości funkcji celu, ograniczenie przyszłych

uprawnień, utratę zasobu lub aktualizację polityki. Uczenie z kosztu nie jest tym samym co przeżywanie kosztu.

Tę granicę trzeba utrzymać, aby AGE pozostawało programem badawczym, a nie metafizycznym skrótem. Jednocześnie biologiczny koszt przypomina o czymś, czego nie wolno redukować do algorytmu. Ludzkie wartościowanie wyrasta z faktu bycia organizmem posiadającym potrzeby, zależności, podatność na uszkodzenia i skończony czas życia. Damasio, tradycja badań nad interocepcją oraz współczesne koncepcje allostazy dowodzą, że kognicja i regulacja biologiczna są ze sobą głęboko sprzężone.

Dla człowieka "ważność" sytuacji nie zawsze jest zewnętrzną wagą dołączoną do reprezentacji; często jest cielesnym stanem organizmu. Nie dowodzi to niemożliwości sztucznego wartościowania, lecz pokazuje, że maszynowy odpowiednik wymagałby znacznie głębszej analizy niż samo dodanie modułu "emotion". Ucieleśnienie ma także wymiar fenomenologiczny i społeczny. Człowiek nie jest abstrakcyjną inteligencją rozwiązującą problemy. Jest ciałem widzialnym dla innych, podlega starzeniu, potrzebuje snu, wchodzi w relacje zależności, wychowuje dzieci, choruje i umiera. Te ograniczenia organizują nie tylko indywidualną psychikę, ale całe instytucje społeczne: rodzinę, medycynę, ubezpieczenia, dziedziczenie, pracę, prawo karne oraz odpowiedzialność.

Hybrydowa ludzkość jako system relacyjny

Dlatego nawet hipotetyczna sztuczna inteligencja, posiadająca funkcjonalnie podobną zdolność rozumowania, nie stałaby się automatycznie tym samym rodzajem podmiotu społecznego. W tym punkcie antropologia technologii spotyka się z posthumanizmem. Donna Haraway wykorzystała figurę cyborga do podważenia rzekomo naturalnych granic między człowiekiem, maszyną a kulturą, natomiast N. Katherine Hayles krytykowała wizję informacji jako wzorca, który można bez strat oddzielić od ucieleśnienia. Wnioskiem z tych rozważań nie jest ani konieczność ochrony biologicznej „czystości”, ani przekonanie, że ciało stało się nieistotne.

Trafniejsza wydaje się teza, że człowiek zawsze był technologicznie relacyjny, lecz poszczególne technologie różnią się stopniem, w jakim uczestniczą w konstytuowaniu jego działania. Z tej perspektywy Hybrydowa Ludzkość nie rozpoczyna się wraz z pierwszym implantem neuronalnym. Zaczyna się wtedy, gdy indywidualnego działania nie da się już adekwatnie wyjaśnić bez uwzględnienia trwałej infrastruktury sztucznej. Osoba korzystająca z AI, która przechowuje historię projektów, formułuje warianty decyzji, organizuje uwagę, zarządza korespondencją i wykonuje zadania w jej imieniu, funkcjonuje w systemie hybrydowym – nawet jeśli wszystkie urządzenia pozostają poza ciałem. Podział na Soft, Embodied i Deep hybrid obrazuje właśnie to kontinuum: od

asystentów cyfrowych, przez biosensory, aż po BCI. Należy jednak dodać, że fizyczna bliskość mózgu nie jest jedyną miarą głębokości integracji. Zewnętrzna infrastruktura może być poznawczo konstytutywna mimo braku implantu, podczas gdy rozwój BCI zwiększa bezpośredniość sprzężenia i przesuwa granicę regulacyjną.

Wówczas prywatność danych ewoluuje w mental privacy, cyberbezpieczeństwo w neurosecurity, a prawo do kontroli nad urządzeniem może stać się warunkiem integralności funkcjonalnej osoby. Nie oznacza to jeszcze „fuzji świadomości", lecz pokazuje, że granica między tym, co biologiczne a techniczne, staje się prawnie i funkcjonalnie mniej oczywista. Hybrydowa Ludzkość wymaga zatem ochrony nie tylko integralności ciała, ale przede wszystkim integralności procesu poznawczego – rozumianej jako możliwość kontroli nad tym, co jest odczytywane, jak interpretowane i jakie interwencje system może wykonywać autonomicznie. Do tej samej kategorii należy problem relacji z AI. Człowiek nie musi wierzyć w świadomość agenta, aby wytworzyć wobec niego autentyczne przywiązanie; psychologia więzi działa bowiem po stronie użytkownika. Jeśli system pamięta historię, reaguje empatycznie, pozostaje stale dostępny i dostosowuje komunikację, może funkcjonalnie wejść w ekologię społeczną człowieka. Problem ten nie daje się zredukować ani do stwierdzenia „to tylko program", ani do uznania „to już osoba". Realna emocja użytkownika może współistnieć z nieustalonym lub negatywnym statusem fenomenalnym partnera. Jest to jedna z charakterystycznych cech nadchodzącej antropologii: realność relacji po jednej stronie nie gwarantuje ontologicznej symetrii po drugiej. W przypadku wychowania dzieci problem ten staje się jeszcze bardziej złożony, gdyż młody człowiek dopiero buduje swoje modele wzajemności, przyjaźni, autorytetu i prywatności.

Mądrość i policentryzm jako bezpieczniki przed algorytmiczną optymalizacją

AI może stać się niezwykle wartościowym tutorem, trenerem mentalizacji i narzędziem wyrównującym różnice w dostępie do wiedzy. Może jednak stworzyć środowisko nadmiernie responsywne, w którym użytkownik rzadziej napotyka opór właściwy relacjom z autonomicznymi ludźmi. W tym kontekście pojawia się potrzeba rodzicielskiej „powściągliwości" wobec technologicznej pokusy optymalizowania rozwoju. Należy ujmować to jako hipotezę dotyczącą rozwojowego znaczenia frustracji, negocjacji i samoregulacji, a nie jako pochwałę cierpienia. Dziecko potrzebuje bezpieczeństwa, ale potrzebuje również możliwości rozwijania kompetencji poprzez problemy, których system nie rozwiązuje za nie natychmiast. Ta sama zasada odnosi się do kognicji dorosłych. Dobra AI nie powinna maksymalizować własnej niezbędności. Jeśli jej

zadaniem jest augmentacja człowieka, miarą sukcesu powinien być wzrost kompetencji użytkownika, funkcjonujących także poza systemem. Tutor osiąga sukces wtedy, gdy uczeń w pewnej klasie zadań coraz mniej go potrzebuje.

Asystent relacyjny jest wartościowy, jeśli pomaga odbudować więzi z ludźmi, a nie gdy optymalizuje jedynie czas spędzony z agentem. System decyzyjny staje się dojrzały, gdy uczy użytkownika prawidłowo identyfikować moment, w którym należy go zakwestionować. Pojęcie mądrości okazuje się tu niezbędne, ponieważ sama inteligencja jest normatywnie zbyt uboga. Inteligencja instrumentalna zwiększa wprawdzie zdolność osiągania dowolnego celu, lecz nie wskazuje, które cele są godne realizacji, jak rozdzielić korzyści i koszty, jak traktować interes przyszłych pokoleń ani kiedy zrezygnować z optymalizacji jednej wartości, by chronić inną. Filozofia praktyczna – od arystotelesowskiej phronesis po współczesne teorie osądu – pokazuje, że mądrość wymaga uwzględnienia konkretnego kontekstu, konfliktowych wartości oraz konsekwencji wykraczających poza bezpośredni sukces instrumentalny. Dlatego refleksja „poza AGI" powinna obejmować nie tylko Artificial General Experience, ale również problem sztucznej lub hybrydowej praktycznej mądrości.

Nie chodzi tu o postulat stworzenia modułu nazwanego „wisdom". Mądrość może być bowiem cechą architektury społecznej, a nie pojedynczego agenta. Instytucja bywa mądrzejsza od swoich członków, jeśli posiada procedury spowalniające decyzje wysokiego ryzyka, niezależny audyt, możliwość odwołania, ochronę mniejszości i zdolność uczenia się na błędach. Może być jednak mniej mądra od ludzi, którzy ją tworzą, gdy system bodźców premiuje krótkoterminowe wyniki i tłumi niewygodne informacje.

W tym miejscu kluczowa staje się teoria instytucji Elinor Ostrom oraz koncepcja policentrycznego zarządzania. Złożone zasoby wspólne nie zawsze wymagają centralnego regulatora; mogą być zarządzane przez nakładające się poziomy reguł, wspólnot i mechanizmów monitorowania. Przyszła infrastruktura AI nie musi zatem być wyborem między globalnym superagentem a całkowitym chaosem. Możliwe są federacje systemów, lokalne centra wiedzy, wspólne standardy, konkurujące modele oraz publiczne instytucje i prywatne narzędzia połączone procedurami interoperacyjności. Taki policentryzm ogranicza ryzyko wspólnego błędu i koncentracji władzy. Teoria systemów złożonych dowodzi ponadto, że odporność wymaga różnorodności i redundancji. Ekosystem z jedną odmianą roślin może być niezwykle wydajny w stabilnych warunkach, a jednocześnie skrajnie podatny na pojedynczy patogen. Analogicznie społeczeństwo, w którym większość decyzji przechodzi przez podobne modele i dane, może zwiększyć krótkookresową efektywność, lecz stać się podatne na wspólny błąd reprezentacji. Epistemiczny pluralizm nie jest więc wyłącznie liberalną wartością kulturową. Jest mechanizmem bezpieczeństwa systemowego.

Pluralizm perspektyw jako fundament wolności epistemicznej

Gujral określa tę intuicję mianem "multi-biased future", argumentując, że całkowite usunięcie biasów jest niemożliwe i niezbędna jest różnorodność perspektyw. Jak wykazano wcześniej, pojęcie to wymaga jednak korekty. Nie każdy bias zasługuje na ochronę; systematyczny błąd czy dyskryminacja powinny być redukowane. Nie można natomiast wyeliminować perspektywiczności modelowania: każda reprezentacja z konieczności wybiera określone cechy, miary i poziom analizy. Właściwym ideałem jest zatem multi-perspectival future – świat, w którym różne modele rzeczywistości konkurują ze sobą i wzajemnie obnażają własne ograniczenia, a ich wyniki są oceniane na podstawie jakości dowodów. Epistemologia społeczeństwa hybrydowego stanie się jednym z kluczowych pól walki o tę różnorodność.

AI zmienia bowiem nie tylko rezultaty (outputs), ale – jak podkreśla materiał – także dane wejściowe ludzkiej kognicji: to, co człowiek czyta, widzi i uznaje za prawdopodobne, a także sposób, w jaki konstruuje problem. Jest to zmiana głębsza niż tradycyjna propaganda, ponieważ wpływ może zostać wbudowany w codzienny mechanizm pomocniczy, którego użytkownik nie postrzega jako formy perswazji. Model organizujący wiedzę może kształtować przestrzeń możliwych poglądów jeszcze przed momentem świadomego wyboru. Dlatego przyszła wolność słowa wymaga uzupełnienia przez wolność epistemicznego środowiska. Prawo do wypowiedzenia dowolnej opinii nie wystarcza, jeśli człowiek funkcjonuje w otoczeniu, w którym jedna infrastruktura systematycznie determinuje, jakie argumenty są warte uwagi. Nie chodzi o stworzenie systemu pozbawionego filtrów – taki układ jest poznawczo niemożliwy. Chodzi o widzialność tych filtrów, możliwość ich zmiany oraz pluralizm ich właścicieli i metod. Wiedza społeczna wymaga ponadto zachowania genealogii twierdzeń.

Generatywna synteza bywa niezwykle użyteczna, gdyż obniża koszt porównywania tysięcy źródeł. Może jednak zacierać informację o tym, które źródła są pierwotne, które jedynie powtarzają cudze ustalenia i gdzie występuje rzeczywisty spór. Przyszłe systemy nie powinny zatem ograniczać się do prostego podawania cytowań. Powinny pomagać odróżniać konsensus niezależny od konsensusu syntetycznego – sytuacji, w której wiele pozornych potwierdzeń ma wspólne źródło lub zostało wygładzone przez ten sam model. W ten sposób kształtuje się idea epistemicznej metakognicji społeczeństwa.

AI może służyć nie tylko do produkowania twierdzeń, lecz także do mapowania jakości wiedzy: wskazywania obszarów o stabilnym konsensusie, identyfikowania konkurencyjnych teorii, wykazania zależności dowodów, braku replikacji czy przestarzałych źródeł oraz momentów, w

których system działa poza rozpoznanym rozkładem. To właśnie byłaby jedna z najcenniejszych funkcji przyszłego Metamind: nie udzielanie odpowiedzi na wszystkie pytania, lecz zwiększanie zdolności zbiorowości do rozpoznawania granic własnego poznania. Metakognicja ma tutaj znaczenie głębsze niż zwykła introspekcja. Na poziomie jednostki obejmuje ocenę pewności, monitorowanie błędu i decyzję o kontynuowaniu deliberacji. Na poziomie organizacji oznacza wiedzę o słabych procesach, konfliktach źródeł i momentach wymagających innej ekspertyzy. Na poziomie cywilizacji mogłaby oznaczać zdolność automatycznego mapowania obszarów, w których społeczeństwo wykazuje nadmierną pewność względem jakości danych. Jest to funkcjonalnie uchwytna idea, znacznie wcześniejsza i naukowo bezpieczniejsza niż spekulacje o "świadomości ludzkości".

Rozszerzona biografia zamiast transferu świadomości

Takie rozróżnienie pozwala uporządkować pojęcia Metamind, Supersociety i Multi-Substrate Person. Metamind w słabym sensie może oznaczać system łączący wielu ludzi i komponenty sztuczne we wspólną architekturę pamięci, koordynacji i uczenia. Supersociety można natomiast rozumieć jako społeczeństwo, w którym znacząca część zbiorowych funkcji poznawczych jest współrealizowana przez ludzi i AI. Multi-Substrate Person oznaczałaby zaś jeden osobowy podmiot przekraczający substrat biologiczny. Pierwsze dwa pojęcia dają się rozwijać jako modele organizacyjne i społeczne; trzecie pozostaje filozoficzną hipotezą dotyczącą tożsamości i świadomości.

Gujral idzie w tej sprawie dalej, sugerując rozproszoną sprawczość oraz możliwość niezależności tożsamości od jednego biologicznego nośnika. W świetle obecnej nauki należy tę propozycję potraktować jako graniczny eksperyment pojęciowy. Z faktu, że organizmy biologiczne wykorzystują wielopoziomową bioelektryczną regulację, nie wynika możliwość przeniesienia osobowej świadomości do krzemu. Z faktu, że zewnętrzna pamięć może funkcjonalnie należeć do rozszerzonego systemu poznawczego, nie wynika, że serwer staje się częścią jednego pola fenomenalnego. Należy utrzymać zasadę: integracja funkcjonalna nie dowodzi integracji fenomenalnej.

Filozofia tożsamości osobowej dodatkowo komplikuje wizję kopiowania umysłu. Teorie ciągłości psychologicznej wskazują na znaczenie pamięci, intencji i związków przyczynowych między kolejnymi stanami osoby, jednak klasyczny problem rozgałęzienia ujawnia, że identyczność informacji nie rozstrzyga identyczności numerycznej. Jeśli powstaną dwa cyfrowe systemy posiadające dokładnie tę samą historię wspomnień, nie mogą one być jednym i tym samym

podmiotem po rozpoczęciu odrębnych historii. Cyfrowa kopia może zatem być psychologicznie bliskim następcą, lecz nie daje odpowiedzi na pytanie o kontynuację pierwszoosobowego doświadczenia.

W praktyce prawdopodobnie znacznie wcześniej pojawi się zjawisko mniej spektakularne, ale społecznie donioślejsze: ciągłość osoby biologicznej przy rosnącym outsourcingu jej funkcji. Człowiek pozostanie jednym organizmem i jednym prawnym podmiotem, lecz część pamięci, planowania, regulacji i komunikacji będzie stale wykonywana przez niebiologiczne komponenty. Zamiast uploadingu otrzymamy rozszerzoną biografię.

Zamiast przeniesienia świadomości – coraz większą współzależność biologicznego centrum doświadczenia i sztucznego środowiska funkcjonalnego. W takim świecie zmieni charakter również problem dziedziczenia. Po śmierci człowieka może pozostać niezwykle rozbudowana pamięć cyfrowa, zdolna do generowania wypowiedzi zgodnych z jego wcześniejszym stylem i historią. Nie oznacza to przetrwania osoby, lecz tworzy nową klasę reprezentacji społecznej. Rodzina może doświadczać obecności cyfrowej persony, przedsiębiorstwo zachować sposób podejmowania decyzji zmarłego założyciela, a instytucja korzystać z jego zewnętrznej pamięci. Prawo będzie musiało rozróżnić dziedziczenie danych, dziedziczenie modelu zachowania oraz ciągłość osoby, gdyż trzy te kategorie nie są tożsame.

Relacja czasu i tożsamości prowadzi z powrotem do najbardziej wartościowej intuicji AGE: tożsamość człowieka nie jest statycznym profilem, lecz historią transformacji. Człowiek pozostaje sobą nie dlatego, że zachowuje identyczne preferencje, ale ponieważ kolejne zmiany pozostają przyczynowo i narracyjnie związane z wcześniejszym życiem. System personalizujący AI powinien więc respektować prawo człowieka do zmiany własnego modelu.

Pefekcyjne przewidywanie oparte na przeszłości nie może stać się normatywną klatką przyszłości. Pozwala to sformułować pojęcie prawa do nieciągłości predykcyjnej. Człowiek powinien posiadać możliwość działania inaczej, niż wynikałoby z jego profilu. Może przestać lubić rzecz, którą wcześniej lubił, zmienić poglądy, porzucić projekt, zerwać nawyk albo zbudować nową tożsamość. System, który zbyt skutecznie dopasowuje przyszłe środowisko do przeszłych preferencji, może zwiększać wygodę, a jednocześnie utrudniać samotransformację. Częścią autonomii jest zatem prawo do zaskoczenia własnego algorytmicznego modelu.

Tutaj technologia spotyka filozofię egzystencji. Ludzkie życie nie jest wyłącznie procesem wyboru spośród przewidywalnych preferencji; obejmuje redefiniowanie własnych celów. Harry Frankfurt odróżniał pragnienia pierwszego rzędu od refleksyjnych pragnień dotyczących tego, jakich pragnień

chcemy być podmiotami. Charles Taylor i tradycje hermeneutyczne wskazywały z kolei, że człowiek interpretuje siebie w obrębie znaczeń kulturowych.

Ochrona suwerenności poznawczej poprzez pluralizm i ramy konstytucyjne

AI zdolna do niezwykle precyzyjnego przewidywania pierwszorzędnych preferencji wciąż nie rozwiązuje problemu refleksyjnej samotransformacji, w ramach której człowiek pyta sam siebie, czy chce pozostać osobą o takich właśnie preferencjach. Zatem kryterium sprawczości nie może sprowadzać się wyłącznie do zgodności działania systemu z aktualnymi pragnieniami użytkownika. AI może bowiem perfekcyjnie realizować to, czego człowiek chce w tej chwili, a jednocześnie utrudniać mu zmianę tego, czego będzie chciał w przyszłości. Dobra infrastruktura autonomii powinna zatem wspierać zarówno realizację preferencji, jak i krytyczną refleksję nad nimi.

W tym punkcie model AI jako mentora okazuje się cenniejszy niż wizja AI jako doskonałego sługi. Gujral wyraźnie przeciwstawia system rzucający poznawcze wyzwania takiemu, który jedynie zastępuje ludzką wolę. Taka relacja wymaga jednak szczególnej ostrożności aksjologicznej.

Mentor bowiem kształtuje rozwój osoby. Jeśli rolę tę przejmuje system kontrolowany przez korporację, państwo lub inną organizację, nieuniknione staje się pytanie: czyje wartości uczestniczą w formowaniu użytkownika? Nie można zatem programować „dobrego człowieka” w oderwaniu od politycznej debaty nad treścią dobra. Hybrydowa edukacja moralna musi być pluralistyczna, transparentna i podporządkowana prawom podmiotu, a nie paternalistycznej funkcji celu. W tym miejscu teoria alignmentu odsłania swoje społeczne ograniczenia. Twierdzenie, że AI należy „dostosować do wartości człowieka”, jest powierzchowne, gdyż nie istnieje jeden, uniwersalny zbiór takich wartości. Społeczeństwa różnią się w kwestiach religii, polityki, obyczajów czy granic wolności; nawet wewnątrz jednej osoby wartości często pozostają w konflikcie. Alignment systemu społecznego jest więc bliższy konstytucjonalizmowi niż technicznemu dostrojeniu funkcji nagrody. Polega on na organizowaniu niezgody według procedur chroniących podstawowe prawa, a nie na poszukiwaniu jednego wektora preferencji całej ludzkości. To obszar, w którym nauki polityczne wnoszą do rozwoju AI element, którego sama informatyka zapewnić nie jest w stanie.

Liberalna demokracja nie dąży do tego, by wszyscy obywatele mieli ten sam cel; tworzy raczej procedury pozwalające ludziom o różnych dążeniach współistnieć i podejmować wspólne decyzje. Konstytucja ogranicza przy tym władzę większości. Podobnie przyszła Supersociety nie powinna optymalizować jednej funkcji globalnego dobra, lecz stanowić architekturę zarządzania

pluralizmem, w której część wartości pozostaje niepodlegająca agregacji. Prawo człowieka jest w takim systemie czymś więcej niż tylko ograniczeniem efektywności – jest strukturalną granicą optymalizacji. System może wyliczyć, że naruszenie prywatności niewielkiej grupy zwiększyłoby dobrobyt większości, jednak porządek konstytucyjny może taką decyzję kategorycznie zabronić. Oznacza to, że mądrość instytucjonalna nie polega na maksymalizacji sumarycznej funkcji użyteczności, lecz obejmuje reguły wyłączające pewne dobra z prostego rachunku kosztów i korzyści.

Problem ten nabiera szczególnego znaczenia wraz ze wzrostem zdolności predykcyjnych AI. Im więcej system potrafi przewidzieć, tym silniejsza staje się pokusa technokratycznego przesunięcia od polityki ku optymalizacji. Jeśli model potrafi oszacować konsekwencje każdego wariantu, może powstać złudzenie, że decyzja polityczna sprowadza się już tylko do wyboru rozwiązania o najwyższej wartości oczekiwanej. Tymczasem modele nie określają same z siebie, które konsekwencje powinny być ważone, czyje koszty są dopuszczalne i jaki rozkład ryzyka jest sprawiedliwy.

> Ostatecznie największym zagrożeniem nie jest powstanie świadomego superkomputera, lecz świat, w którym perfekcyjnie dopasowany do nas algorytm staje się tak wygodny, że przestajemy chcieć być kimś innym. Czy w rzeczywistości, która przewiduje każdy nasz ruch i spełnia każdą preferencję, pozostanie jeszcze miejsce na prawo do błędu, zaskoczenia i autentycznej samotransformacji? Prawdziwa wolność w epoce hybrydowej może okazać się nie tyle zdolnością do korzystania z AI, co odwagą do bycia nieprzewidywalnym dla własnego modelu.

Inspiracje

[1] **"The AI Instinct"** Rana Gujral

Rana Gujral (ur. 1976) to amerykański przedsiębiorca, dyrektor i autor specjalizujący się w kognitywnej sztucznej inteligencji, rozpoznawaniu emocji i podejmowaniu decyzji w relacji człowiek-komputer. Jest powszechnie uznawany za dyrektora generalnego Behavioral Signals, firmy technologicznej dla przedsiębiorstw, która opracowuje oprogramowanie do głębokiego uczenia się do analizy intencji, emocji i sygnałów behawioralnych na podstawie danych głosowych. Przed objęciem stanowiska dyrektora Behavioral Signals, Gujral założył platformę chmurową dla przedsiębiorstw TiZE, którą przejęła firma Alchemy, a także pełnił funkcje kierownicze i odpowiadał za transformację produktów w firmach Cricut i Logitech. W swoich pracach, w tym w książce „Instynkt AI: Przyszłość sztucznej inteligencji i podejmowania decyzji przez człowieka", Gujral analizuje ewolucję sztucznej inteligencji w kierunku sztucznej inteligencji ogólnej (AGE), omawiając, jak hybrydowe poznanie i nowe technologie sztucznej inteligencji zmieniają ludzki osąd, sprawczość i przywództwo organizacyjne.

Rana Gujral *(born 1976) is an American entrepreneur, executive, and author specializing in cognitive artificial intelligence, emotion recognition, and human-computer decision-making. He is widely recognized as the chief executive officer of Behavioral Signals, an enterprise technology company that develops deep-learning software to analyze intent, emotion, and behavioral cues from vocal data. Prior to leading Behavioral Signals, Gujral founded the cloud enterprise platform TiZE, which was acquired by Alchemy, and held leadership and product turnaround roles at Cricut and Logitech. In his work, including his book The AI Instinct: The Future of AI and Human Decision-Making, Gujral explores the evolution of machine intelligence toward Artificial General Experience (AGE), addressing how hybrid cognition and emerging AI technologies reshape human judgment, agency, and organizational leadership.*

Behavioral Signals

Literatura uzupełniająca

Książki

Książki autorów przywoływanych i literatura pokrewna:

[1] **"Progress"**

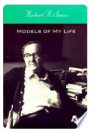
N. Katherine Hayles



[2] "Organizations"

Herbert Simon

John Wiley & Sons | ISBN: 9780631186311



[3] "Models of my life"

Herbert Simon

MIT Press | ISBN: 9780262326582



[4] "Human Problem Solving"

Herbert Simon

Prentice Hall



[5] "Design for a brain"

W. Ross Ashby

ISBN: 9781614277569



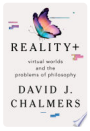
[6] "Design for a brain; the origin of adaptive behavior"

W. Ross Ashby

ISBN: 9781614277569

[7] "An introduction to cybernetics"

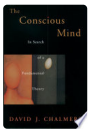
W. Ross Ashby



[8] "Reality+"

David Chalmers

W. W. Norton & Company | ISBN: 9780393635812



[9] "The Conscious Mind"

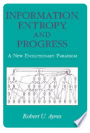
David Chalmers

Oxford University Press | ISBN: 9780199839353

[10] "The rational and social Foundations of music"

Max Weber

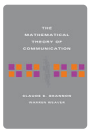
ISBN: 9780758136602



[11] "information entropy"

Claude Shannon

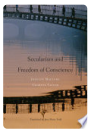
Springer Science & Business Media | ISBN: 9780883189115



[12] "The mathematical theory of communication"

Claude Shannon

University of Illinois Press | ISBN: 9780252098031



[13] "Secularism and Freedom of Conscience"

Charles Taylor

Harvard University Press | ISBN: 9780674265721



[14] "The Algorithmic Alchemist: Forging Sentience in the Silicon Crucible"

Adrian Cole

nick creighton



[15] "The Manush AI Blueprint: AGI Research, Humanoid Robotics, and the Geometry of Consciousness"

ABHIJEET SARKAR

2026 | ABHIJEET SARKAR | ISBN: 9798235645370



[16] "How AGI Could Be Humanity's Last Invention"

Ethan K Ben

2025 | Independently Published | ISBN: 9798297369290

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Communication Theory of Secrecy Systems*"

Claude E. Shannon

1949 | Bell System Technical Journal | DOI: 10.1002/j.1538-7305.1949.tb00928.x

Źródło

[2] "Observations and Comments on the Organization Studies of the Systems Research Laboratory"

Herbert Simon

1952 | RAND Corporation eBooks | DOI: 10.7249/rm922

Źródło

[3] **"The Cost of Paying Attention: Cognitive Resource Scarcity and Investor Activity Around FDA Announcements"**

Herbert Simon, Doron Kliger, Ofir Levi, Tiran Rothman

2014

Źródło

[4] **"Design for a Brain"**

W. Ross Ashby

1960 | Springer eBooks | DOI: 10.1007/978-94-015-1320-3

Źródło

[5] **"Design for an Intelligence-Amplifier"**

W. Ross Ashby

1956 | Princeton University Press eBooks | DOI: 10.1515/9781400882618-011

Źródło

[6] **"The state-determined system."**

W. Ross Ashby

2008 | Chapman & Hall eBooks | DOI: 10.1037/11592-018

Źródło

[7] **"Adaptation in iterated and serial systems."**

W. Ross Ashby

2008 | Chapman & Hall eBooks | DOI: 10.1037/11592-014

Źródło

[8] **"Whatever next? Predictive brains, situated agents, and the future of cognitive science"**

Andy Clark

2013 | Behavioral and Brain Sciences | DOI: 10.1017/s0140525x12000477

Źródło

[9] **"The Extended Mind"**

Andy Clark, David J. Chalmers

1998 | Analysis | DOI: 10.1093/analys/58.1.7

Źródło

[10] "An embodied cognitive science?"

Andy Clark, Andy Clark

1999 | Trends in Cognitive Sciences | DOI: 10.1016/s1364-6613(99)01361-3

Źródło

[11] "Mindware: An Introduction to the Philosophy of Cognitive Science"

Andy Clark

2000

Źródło

[12] "Facing Up to the Problem of Consciousness"

David J. Chalmers

1996 | The MIT Press eBooks | DOI: 10.7551/mitpress/6860.003.0003

Źródło

[13] "The Character of Consciousness"

David J. Chalmers

2010 | Oxford University Press eBooks | DOI: 10.1093/acprof:oso/9780195311105.001.0001

Źródło

[14] "Metametaphysics : new essays on the foundations of ontology"

David J. Chalmers, David Manley, Ryan Wasserman

2009

Źródło

[15] "Thinking About Brain and Consciousness"

Antonio Damasio

2011 | Research and Perspectives in Neurosciences | DOI: 10.1007/978-3-642-18015-6_3

Źródło

[16] "4. Music Education and Child Development"

Assal Habibi, Hanna Damasio, Antonio Damasio

2021 | Classical Music | DOI: 10.11647/obp.0242.04

Źródło

[17] "The neural substrates of cognitive empathy"

Stephanie D. Preston, A. Bechara, Hanna Damasio, T. Grabowski, R. Stansfield

2007 | Social Neuroscience | DOI: 10.1080/17470910701376902

Źródło

[18] "Sensing, feeling and consciousness"

Antonio Damasio, Hanna Damasio

2024 | Philosophical Transactions of the Royal Society of London. Biological Sciences | DOI: 10.1098/rstb.2023.0243

Źródło

[19] "Ethics for Cognitive Assemblages: Who's in Charge Here?"

N. Hayles

2022 | Palgrave Handbook of Critical Posthumanism | DOI: 10.1007/978-3-030-42681-1_11-1

Źródło

[20] "Governing the Commons: The Evolution of Institutions for Collective Action"

Barry C. Field, Элинор Остром

1992 | Land Economics | DOI: 10.2307/3146384

Źródło

[21] "Energetic cost of carrying loads: have African women discovered an economic way?"

Geoffrey M. O. Maloiy, Norman C. Heglund, Laura M. Prager, G Cavagna, Charles Ray Taylor

1986 | Nature | DOI: 10.1038/319668a0

Źródło

Artykuły naukowe autora:

[22] "The Nature of Human Cognition and AI's Approach to Mimicking It"

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch02

Źródło

[23] "AI in the Real World: Applications and Use Cases"

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch03

Źródło

[24] "The AI Instinct"

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.fmatter

Źródło

[25] "The Emergence of Artificial General Intelligence (AGI)"
RANA GUJRAL, KAI-FU LEE
2026 | DOI: 10.1002/9781394392155.ch04
Źródło

[26] "Introduction: Our Definition of “There” Is Wrong"
RANA GUJRAL, KAI-FU LEE
2026 | DOI: 10.1002/9781394392155.ch01
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: e390a1c0-04a8-470b-b9d9-a542b0beb244