Architektura hybrydowej kognicji: projektowanie odporności i odpowiedzialności w świecie po AGI w świetle The AI Instinct

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Tekst analizuje konieczność budowy odpornych systemów człowiek-AI, odrzucając dążenie do całkowitej eliminacji tarcia poznawczego. Autor argumentuje, że proaktywne tarcie i redundancja są kluczowe dla bezpieczeństwa globalnego, zapobiegając katastrofalnym błędom optymalizacji wąskich funkcji celu. Kluczowym elementem jest koncepcja 'responsibility by design', która przenosi odpowiedzialność z pojedynczego operatora na poziom architektury systemowej. Postuluje się wprowadzenie mechanizmów contestability (możliwości zakwestionowania decyzji AI) oraz wielowarstwowych zabezpieczeń (defence in depth). W świecie post-AGI governance staje się integralną częścią technologii, wymagając nowych ram prawnych dla relacji funkcjonalnych w systemach hybrydowych, gdzie odpowiedzialność jest ściśle powiązana z realną kontrolą i możliwością przewidywania zdarzeń.

The text analyzes the necessity of building resilient human-AI systems, rejecting the pursuit of total elimination of cognitive friction. The author argues that proactive friction and redundancy are crucial for global safety, preventing catastrophic errors caused by the optimization of narrow objective functions. A key element is the concept of 'responsibility by design,' which shifts responsibility from a single operator to the level of system architecture. It advocates for the introduction of contestability mechanisms (the ability to challenge AI decisions) and multi-layered safeguards (defense in depth). In a post-AGI world, governance becomes an integral part of technology, requiring new legal frameworks for functional relationships in hybrid systems, where responsibility is closely linked to real control and predictability.

Artykuł stanowi krytyczną analizę przejścia ludzkości w erę hybrydowej inteligencji, odrzucając wizję technologicznego determinizmu na rzecz koncepcji świadomego projektu instytucjonalnego. Autor stawia tezę, że kluczowym wyzwaniem nie jest samo osiągnięcie AGI czy zwiększanie mocy obliczeniowej modeli, lecz stworzenie architektury governance, która zapobiegnie atrofii ludzkich kompetencji i utracie sprawczości. Tekst argumentuje, że bezpieczeństwo systemowe wymaga wprowadzenia „celowego tarcia" – mechanizmów kontroli i weryfikacji, które choć spowalniają procesy, chronią przed katastrofalnymi błędami optymalizacji wąskich celów. W centrum rozważań znajduje się postulat przejścia od badania samej sztucznej inteligencji ku „nauce o kognicji hybrydowej" (science of hybrid cognition). Ma ona na celu zdefiniowanie nowych ram odpowiedzialności, higieny poznawczej oraz praw człowieka w świecie, gdzie granica między ludzkim a maszynowym przetwarzaniem informacji ulega zatarciu, a godność jednostki zostaje oddzielona od jej produktywności w benchmarkach AI.

SŁOWA KLUCZOWE

architektura hybrydowej kognicji | odporność systemowa | AGI | The AI Instinct | alignment layer | responsibility by design | moral crumple zone | contestability | hybrydowa inteligencja | infrastructural transition | higiena kognitywna | science of hybrid cognition | normative governability | epistemiczna redundancja

Odporność systemowa wymaga celowego tarcia i odpowiedzialności projektowej

Ekonomia dobrobytu i teoria wyboru społecznego od dawna wskazują na trudność agregowania heterogenicznych preferencji. Twierdzenie Arrowa ujawnia ograniczenia idealnego mechanizmu agregacji rankingów społecznych przy określonym zestawie wymagań, natomiast Sen rozszerzył tę debatę o prawa, wolność i zdolności jednostek, dowodząc, że dobrobytu nie można sprowadzić do jednej miary zasobów czy preferencji. W świecie AI te klasyczne problemy nie znikają; stają się wręcz bardziej widoczne, gdyż system może z ogromną precyzją optymalizować źle zdefiniowaną funkcję celu. Dlatego po AGI potrzebna będzie nie tylko lepsza technologia, ale przede wszystkim

lepsza teoria tego, co wolno jej optymalizować. Nie jest to już problem wyłącznie techniczny – wymaga on zaangażowania ekonomii politycznej, prawa, etyki, teorii sprawiedliwości, psychologii oraz demokratycznej legitymizacji. System może być technicznie aligned z jawnie określonym celem, a jednocześnie społecznie szkodliwy, jeśli sam cel jest zbyt wąski.

Podobną lekcję daje ekologia. Maksymalizowanie pojedynczego parametru produktywności często prowadzi do degradacji odporności całego ekosystemu. Monokultura rolnicza zwiększa wydajność konkretnej uprawy w krótkim okresie, lecz jednocześnie obniża odporność na patogeny i zmiany klimatyczne. Analogicznie społeczeństwo maksymalizujące szybkość i efektywność poznawczą może tracić redundancję, różnorodność oraz powolne, lecz krytyczne instytucje niezbędne podczas kryzysu. Nieefektywność lokalna może czasami być ceną odporności globalnej.

To kluczowy argument przeciwko usuwaniu każdego „tarcia" z systemu człowiek-AI. Tarcie bywa kosztem, ale pełni również istotną funkcję. Procedura odwoławcza spowalnia decyzję, lecz chroni przed błędem; niezależna recenzja opóźnia publikację, ale zwiększa szansę na wykrycie usterki; drugi podpis przy operacji wysokiego ryzyka jest nieefektywny w większości przypadków, jednak może zapobiec katastrofie.

Proaktywne tarcie należy zatem interpretować jako inżynierię odporności, a nie romantyczną pochwałę ludzkiej powolności. Teoria niezawodności wprowadza tu zasadę defence in depth: żaden pojedynczy mechanizm bezpieczeństwa nie powinien być uznany za niezawodny. System krytyczny potrzebuje warstw, których błędy nie są całkowicie skorelowane. W kontekście AI oznacza to połączenie ograniczeń technicznych, niezależnych modeli, nadzoru człowieka, audytu, logów, prawa do zatrzymania, procedur odwoławczych i instytucjonalnej odpowiedzialności. Im bardziej agentowa staje się infrastruktura, tym mniej zasadne jest przekonanie, że bezpieczeństwo zapewni pojedynczy „alignment layer". Z tego punktu widzenia odpowiedzialność staje się własnością architektury. Moral crumple zone pojawia się wtedy, gdy organizacja posiada technologicznie rozproszoną kontrolę, lecz po wystąpieniu błędu koncentruje winę na człowieku znajdującym się najbliżej ostatniego przycisku. Gujral uznaje taki stan za problem etyczny systemów hybrydowych. Przyszłe systemy powinny wiązać odpowiedzialność z wiedzą, kontrolą, możliwością przewidywania i realną zdolnością zapobiegania zdarzeniom. Operator, integrator, producent, właściciel danych i organizacja wdrażająca mogą posiadać różne stopnie tej odpowiedzialności.

Nie oznacza to rozmycia winy. Wręcz przeciwnie – responsibility by design wymaga ustalenia odpowiedzialności przed zdarzeniem, a nie dopiero po nim. Jeśli nie wiadomo, kto odpowiada za kalibrację, dane czy uprawnienia agenta i kto ma prawo zatrzymać system, architektura jest niedojrzała niezależnie od jakości modelu. Governance jest zatem integralną częścią technologii. Prawo przyszłości będzie musiało regulować nie tylko „AI" jako przedmiot, ale relacje funkcjonalne

w systemach hybrydowych. Tradycyjne kategorie producenta, użytkownika i produktu mogą okazać się zbyt proste, gdy system stale się aktualizuje, działa agentowo i współuczy z organizacją. Potrzebne będą pojęcia stopniowalnej autonomii, obowiązku eskalacji, audytowalnej pamięci, odpowiedzialności za aktualizacje oraz prawa do funkcjonalnej ciągłości. Na poziomie epistemicznym analogiczną funkcję powinno pełnić prawo do zakwestionowania syntetycznej wiedzy. Jeżeli AI uczestniczy w decyzji o leczeniu, kredycie, uprawnieniu administracyjnym lub ocenie człowieka, sama możliwość poznania finalnej rekomendacji nie wystarcza. Niezbędna jest możliwość przedstawienia informacji pominiętej przez system oraz uzyskania rzeczywistej rewizji. Contestability jest więc jednocześnie prawem proceduralnym i mechanizmem uczenia instytucji. Tutaj ponownie spotykamy AGE.

Organizacja, która koryguje pojedynczą decyzję, lecz nie zmienia własnego procesu, nie nauczyła się w sensie systemowym. Organizacja, która po incydencie modyfikuje dane, procedury, uprawnienia i sposób szkolenia, przechodzi historyczną transformację instytucjonalną. Można ją uznać za analog doświadczenia bez przypisywania organizacji fenomenalnej świadomości. Ta słaba wersja AGE może być jednym z najważniejszych praktycznych kierunków rozwoju AI: budowaniem systemów, które nie tylko rejestrują błąd, ale zmieniają własną architekturę postępowania po zrozumieniu jego przyczyn.

Hybrydowa inteligencja jako system organizacji, a nie nowy organizm

Podobne zjawisko można dostrzec w ewolucji biologicznej, choć analogię tę należy stosować ostrożnie. Ewolucja nie posiada intencji ani centralnego podmiotu, lecz utrwala konsekwencje wcześniejszych interakcji w strukturze populacji. Kultura działa podobnie, choć w znacznie szybszej skali: procedury, normy i technologie zapisują rozwiązania dawnych problemów. AI może stać się dodatkową warstwą aktywnego dziedziczenia kulturowego, zdolną nie tylko do przechowywania reguł, lecz do rekonstrukcji warunków, które doprowadziły do ich powstania. Teoria ewolucyjnych "major transitions", kojarzona z Johnem Maynardem Smithem i Eörsem Szathmárym, dostarcza tutaj użytecznej, choć nie rozstrzygającej analogii. W historii życia dochodziło bowiem do przejść, podczas których wcześniej autonomiczne jednostki tworzyły wyższy poziom organizacji – jak komórki w organizmach wielokomórkowych. Warunkiem stabilności takich procesów były mechanizmy koordynacji, ograniczania konfliktów i transmisji informacji.

Nie ma podstaw, by ogłaszać, że Hybrydowa Ludzkość jest kolejnym biologicznym major transition. Można jednak analitycznie zapytać, czy rosnąca integracja ludzi, AI i instytucji tworzy nowe

jednostki funkcjonalne, posiadające trwałość, pamięć i cele na wyższym poziomie organizacji. Właśnie w tym sensie Metamind jest wartościowszy jako pytanie niż jako odpowiedź. Jeżeli przedsiębiorstwo, laboratorium czy instytucja dysponuje długotrwałą pamięcią, autonomicznie rozdziela zasoby, monitoruje własne błędy i zachowuje ciągłość działania mimo wymiany ludzi i modeli, można zasadnie pytać o funkcjonalną indywidualność takiego systemu.

Nie trzeba jednak od razu przypisywać mu świadomości. Nauka powinna najpierw zbadać granice systemu: jego zdolność uczenia się, stabilność celów, adaptacyjność oraz stopień interdependencji komponentów. Prowadzi to do istotnego rozróżnienia między inteligencją zbiorową, sprawczością zbiorową a podmiotowością zbiorową. Pierwsza oznacza zdolność grupy do rozwiązywania problemów. Druga wymaga mechanizmu działania w imieniu całości. Trzecia, w mocnym sensie, sugerowałaby istnienie jednego doświadczenia lub jednego podmiotu moralnego. Społeczeństwa i organizacje wyraźnie realizują pierwszą z tych cech, częściowo drugą, natomiast trzecia pozostaje problematyczna. Future studies powinny unikać przenoszenia terminologii z jednego poziomu na drugi. Taki rygor chroni również przed niebezpieczną metaforą społeczeństwa jako organizmu. W biologicznym organizmie pojedyncza komórka nie posiada praw przeciwko całości. W demokratycznym społeczeństwie jednostka posiada prawa właśnie po to, aby interes zbiorowości nie mógł całkowicie ją pochłonąć. Jeżeli Supersociety stanie się niezwykle efektywnym systemem poznawczym, nie powinna z tego wynikać normatywna zasada, że ludzie są jedynie jego komponentami. Zdolność całości do myślenia nie anuluje moralnego statusu części. To jedna z najważniejszych granic przyszłego posthumanizmu.

Krytyka autonomicznego liberalnego podmiotu jest trafna o tyle, że człowiek rzeczywiście kształtuje się poprzez zależności i relacje. Nie wolno jednak z tej relacyjności wyprowadzać prawa sieci do pełnego podporządkowania jednostki. Hybrydowa Ludzkość potrzebuje antropologii relacyjnej połączonej z prawami chroniącymi przed dominacją. Człowiek jest częścią większych systemów, ale nie staje się przez to wyłącznie ich zasobem. Tu pojawia się również kwestia godności.

Godność człowieka nie zależy od przewagi nad AI

Jeśli AI sprawia, że część ludzkich kompetencji traci ekonomiczną wyjątkowość, nie oznacza to spadku wartości człowieka. Godność nie może być uzależniona od przewagi nad maszyną w szachach, języku, matematyce czy kodowaniu; taka koncepcja byłaby skrajnie krucha, gdyż każdy postęp technologiczny uszczuplałby zakres ludzkiej wartości. W nowoczesnych systemach praw człowieka godność przypisana jest osobie, a nie jej pozycji w benchmarku produktywności. Właśnie dlatego pytanie o „wyjątkowość człowieka” wymaga przeformułowania. Nie musimy udowadniać

istnienia jednej kompetencji, której maszyna nigdy nie osiągnie. Znacznie ważniejsze jest rozróżnienie deskryptywnej wyjątkowości i normatywnego statusu. Człowiek może przestać być jedynym systemem zdolnym do określonego rodzaju twórczości, a nadal posiadać prawa i interesy wynikające z bycia ucieleśnionym, społecznym i – zgodnie z naszą najlepszą wiedzą – świadomym podmiotem, zdolnym do cierpienia oraz prowadzenia własnego życia. Jeśli pewnego dnia pojawią się mocne dowody na sentience systemów sztucznych, teoria statusu moralnego będzie musiała zostać rozszerzona. Nie można wykluczyć takiej możliwości jedynie ze względu na substrat, ani przypisywać jej wyłącznie na podstawie płynnego języka. Moralna ostrożność wymaga podejścia probabilistycznego. To właśnie dlatego w poprzednich częściach rozdzielono pojęcia intelligence, agency, self-model, consciousness i sentience – w przyszłości mogą one występować w różnych konfiguracjach.

Możliwe jest zatem, że wysoce agentowa AI pozostanie nieświadoma, podczas gdy bardzo prosty organizm biologiczny będzie posiadał minimalny zakres doświadczenia. Możliwe również, że odpowiednia architektura sztuczna uzyska cechy świadomości, których dziś nie potrafimy wiarygodnie testować. Normatywnie kluczowe jest jednak uniknięcie automatycznego przejścia od mocy do statusu moralnego. Bardzo kompetentny system nie zasługuje na prawa dlatego, że jest kompetentny; jeśli kiedykolwiek uzyska interes moralny, będzie to zależało od innych właściwości, takich jak zdolność do doświadczania dobra i szkody. Takie ujęcie pozwala uporządkować klasyczne pytanie o ASI. Superinteligencja jest przede wszystkim relatywnym pojęciem zdolności: oznacza przewagę nad ludzką kompetencją w szerokim zakresie, lecz nie przesądza o świadomości, wartościach, mądrości ani statusie osobowym.

Zatem wizja superinteligencji jako fuzji człowieka i AI jest interesującym scenariuszem, ale nie rozwiązuje automatycznie problemu bezpieczeństwa. Jeśli potężny system staje się hybrydą, problem alignment przenosi się do jego wnętrza: pytanie brzmi, które grupy ludzkie posiadają władzę, jakie wartości zakodowano w instytucjach, kto kontroluje modele i jak rozwiązywane są konflikty. Nie istnieje bowiem jedna „ludzkość” stojąca po jednej stronie i jedna „AI” po drugiej. Istnieją państwa, przedsiębiorstwa, klasy społeczne, wspólnoty, profesje, modele open source, prywatne infrastruktury, agenci osobowi i systemy wojskowe.

Przyszłość AI to problem projektowy i instytucjonalny, a nie technologiczny determinizm

Przyszłość będzie prawdopodobnie ekologią inteligencji, a nie pojedynczą fuzją. Relacje w tym układzie będą zróżnicowane: jedne symbiotyczne, inne konkurencyjne; część oparta na dominacji

infrastrukturalnej, a część na współpracy federacyjnej. Metafora ekologii jest trafniejsza niż wizja jednego globalnego superumysłu, gdyż lepiej oddaje istnienie nisz, przepływy zasobów i proces koewolucji. Technologia zmienia środowisko selekcji ludzkich kompetencji, podczas gdy ludzie modyfikują sposób korzystania z technologii. Instytucje adaptują procedury do nowych możliwości, a te z kolei kształtują następne generacje modeli. Koewolucja nie oznacza jednak harmonii; może obejmować konflikty, relacje pasożytnicze, uzależnienia i wzajemną adaptację. Przeciwstawienie „wyścigu z maszynami" i „ewolucji równoległej" należy zatem zachować jako korektę uproszczonej narracji, choć nie stanowi ono gwarancji pozytywnego rezultatu. Koewolucja może prowadzić zarówno do augmentacji, jak i atrofii kompetencji; do demokratyzacji wiedzy albo koncentracji władzy; do pluralizmu bądź homogenizacji kultury.

Wynik ten nie jest technologicznie zdeterminowany. Technologia tworzy przestrzeń możliwości, lecz to instytucje określają prawdopodobieństwo ich realizacji. Ten sam model może być nauczycielem rozwijającym kompetencje lub automatem przejmującym wysiłek ucznia. System rekomendacyjny może poszerzać dostęp do wiedzy albo zamykać człowieka w personalizowanej bańce. BCI może przywracać komunikację osobie sparaliżowanej, ale w przyszłości zostać włączone w systemy nadzoru.

Społeczny sens technologii nie istnieje niezależnie od jej instytucjonalnego umiejscowienia, co od dawna podkreśla socjologia technologii. Artefakty nie trafiają do próżni; ich znaczenie kształtuje się poprzez praktyki, normy, interesy i systemy władzy. AI jest przypadkiem szczególnym, gdyż sama zaczyna organizować te praktyki. Nie tylko zostaje społecznie skonstruowana, ale zwrotnie współkonstruuje środowisko społeczne, w którym powstają kolejne normy. Ta pętla zwrotna nadaje kluczowe znaczenie obecnym decyzjom projektowym. Domyślne ustawienia mogą stać się zwyczajem, zwyczaj standardem, standard infrastrukturą, a infrastruktura ostatecznym warunkiem wszystkich przyszłych wyborów. Jest to klasyczna path dependence. Niewielka decyzja na początku rozwoju technologii może, poprzez efekty sieciowe i koszty zmiany, prowadzić do trwałego układu instytucjonalnego. Źródłowy „drift" można więc rozumieć jako instytucjonalne osadzanie się decyzji, których pojedynczo nie postrzegamy jako ustrojowych. Właśnie dlatego pytanie „co poza AGI?" jest w istocie pytaniem o doktrynę projektowania przejścia.

Techniczna roadmapa zwiększania możliwości modeli jest niewystarczająca. Potrzebne są równoległe strategie zachowania ludzkich kompetencji, praw, różnorodności epistemicznej, interoperacyjności, odpowiedzialności oraz instytucjonalnej zdolności do korekty. Rozwój capability bez rozwoju governance tworzy asymetrię: system staje się coraz sprawniejszy, podczas gdy społeczeństwo pozostaje nieprzygotowane do zarządzania jego konsekwencjami. W tym kontekście można zaproponować wielowymiarową teorię rozwoju po AGI.

Pierwszym wymiarem jest capability – szerokość i poziom kompetencji. Drugim adaptacyjna historyczność, czyli funkcjonalny sens AGE: trwała pamięć, uczenie się z konsekwencji i transformacja polityki działania. Trzeci to agency, rozumiana jako zdolność inicjowania i prowadzenia wieloetapowych działań. Czwarty dotyczy metakognicji – kalibracji, rozpoznawania własnych ograniczeń i umiejętności eskalacji. Piąty obejmuje embeddedness, czyli integrację systemu ze środowiskiem społecznym, fizycznym i instytucjonalnym. Szósty to normative governability – zdolność włączenia systemu w strukturę odpowiedzialności, praw, contestability i kontroli. Wymiary te nie stanowią jednego indeksu i nie powinny być redukowane do akronimu. Poza nimi leży problem doświadczenia fenomenalnego, którego nie wolno mechanicznie włączać do tej samej skali. Świadomość może być niezależnym wymiarem ontologicznym. Możliwe, że określony poziom integracji funkcjonalnej okaże się wystarczający; możliwe, że nie. Dopóki spór naukowy trwa, rozsądniej jest budować prawo i bezpieczeństwo w oparciu o właściwości obserwowalne, pozostawiając przestrzeń na przyszłą rewizję statusu moralnego.

Taka wielowymiarowość wyjaśnia również, dlaczego AGI może stracić na znaczeniu jeszcze przed swoim definitywnym osiągnięciem. Jeśli społeczeństwa zaczną korzystać z systemów szerokich (choć niespełniających jednej z definicji AGI), a zarazem silnie agentowych, posiadających pamięć i zintegrowanych z instytucjami, transformacja społeczna może być już głęboka. Granice klasyfikacyjne pozostaną domeną sporów akademickich, podczas gdy gospodarka i polityka będą funkcjonować w warunkach hybrydowej kognicji.

Przejście infrastrukturalne wymaga nowej higieny kognitywnej

Historia technologii zna podobne przypadki. Rewolucja przemysłowa nie rozpoczęła się nagle od wynalezienia jednej „maszyny ogólnego przeznaczenia". Była raczej efektem akumulacji zmian w energetyce, transporcie, organizacji produkcji, finansach, urbanistyce, rynku pracy i administracji. Elektryfikacja również nie sprowadzała się do jednego urządzenia, lecz stanowiła gruntowną przebudowę infrastruktury. AI może podążać podobną drogą: jej najgłębszym skutkiem może nie być pojedynczy model, lecz reorganizacja instytucji wokół powszechnej zdolności maszynowego przetwarzania, generowania treści i działania. Z tej perspektywy singularity okazuje się mniej użyteczną metaforą niż infrastructural transition. Zamiast jednego momentu, po którym wszystko staje się nieprzewidywalne, możemy obserwować serię zmian progowych w różnych domenach.

Edukacja osiąga własny próg, gdy AI staje się domyślnym partnerem uczenia. Prawo inny, gdy większość dokumentacji i procesów wyszukiwania przechodzi przez modele. Nauka kolejny, gdy

hipotezy i eksperymenty są częściowo generowane automatycznie. Administracja następny, gdy agentowa infrastruktura uczestniczy w realizacji polityk publicznych.

Hybrydowa Ludzkość powstaje poprzez nakładanie się tych progów, a nie w konkretnej dacie. Właśnie dlatego społeczna percepcja może długo nie dostrzegać skali transformacji. Każda lokalna zmiana jawi się jako wygodne usprawnienie: lepszy edytor, szybsze wyszukiwanie, inteligentniejszy kalendarz, sprawniejsza diagnoza czy automatyczna synteza. Dopiero ich integracja ujawnia, że człowiek coraz częściej funkcjonuje poprzez środowisko, które nie tylko wykonuje instrukcje, lecz rekonstruuje intencje, przewiduje potrzeby i inicjuje działania. Źródłowy „drift" nie jest zatem spektakularną utratą władzy na rzecz maszyny, lecz możliwością normalizacji delegacji. Odpowiedzią na to nie jest jednak całkowite zatrzymanie tego procesu.

Cywilizacja istnieje dzięki podziałowi pracy i outsourcingowi poznawczemu. Rozsądna reakcja polega na odróżnianiu funkcji, których delegacja zwiększa sprawczość, od tych, których trwałe przekazanie może ją osłabić. Automatyzacja rachunków zwiększyła zdolność ludzi do przeprowadzania bardziej zaawansowanych analiz. Z kolei automatyzacja nawigacji może prowadzić do atrofii części zdolności przestrzennych. AI może analogicznie uwalniać człowieka od mechanicznego wyszukiwania, a jednocześnie – jeśli zostanie źle zaprojektowana – osłabiać jego umiejętność rozpoznawania jakości argumentu. Tak dochodzimy do pojęcia higieny kognitywnej epoki AI. Nie powinna ona przyjmować postaci ascetycznego zakazu korzystania z technologii, lecz polegać na świadomym różnicowaniu trybów jej użycia. Niektóre zadania można delegować całkowicie, inne powinny być wykonywane wspólnie, a jeszcze inne warto czasem realizować bez AI właśnie po to, aby utrzymać kompetencje.

Wysokość tego „minimum samodzielności" będzie zależeć od zawodu, poziomu ryzyka i funkcji społecznej. Edukacja stanie się miejscem, w którym społeczeństwo będzie negocjować te granice najintensywniej. Szkoła nie może konkurować z AI w zakresie ilości natychmiast dostępnych informacji; jej rolą stanie się rozwijanie modeli pojęciowych, zdolności rozumowania, epistemicznej kontroli, umiejętności definiowania problemów oraz kompetencji współdziałania z systemami sztucznymi. Nie oznacza to porzucenia wiedzy faktograficznej, gdyż bez wiedzy domenowej człowiek nie potrafi sprawnie oceniać odpowiedzi AI. „Umiejętność zadawania pytań" bez aparatu wiedzy jest równie niewystarczająca, co pamięciowe uczenie bez rozumienia. Przyszła edukacja powinna zatem kształcić dwa rodzaje kompetencji równocześnie: zdolność samodzielnej konstrukcji modeli świata oraz umiejętność strategicznego rozszerzania ich przez AI. Nie chodzi tu o szkolenie z prompt engineering.

Jest to rozwój rozszerzonej funkcji wykonawczej: wiedza, kiedy delegować, kiedy zachować niezależność, jak interpretować niepewność i kiedy niezbędne jest źródło pierwotne. Człowiek

przyszłości musi być ekspertem nie tylko w swojej dziedzinie, ale także w architekturze własnego procesu poznawczego. Na poziomie społecznym odpowiednikiem tej kompetencji jest institutional AI literacy. Organizacja musi wiedzieć, gdzie znajdują się jej zależności, które funkcje nie mogą opierać się na jednym modelu, jaki jest plan działania po awarii, które dane należy chronić i gdzie niezbędny jest człowiek posiadający niezależną kompetencję. Wysoki poziom technologiczny bez takiej metawiedzy może zwiększać sprawność, a jednocześnie pogłębiać kruchość systemu. To samo dotyczy państwa demokratycznego.

Wraz z rozwojem AI wzrośnie potencjalna zdolność administracji do analizowania zachowań i przewidywania skutków polityk. Może to poprawić jakość usług publicznych, ale może także zwiększyć asymetrię wiedzy między państwem a obywatelem. Dlatego technologiczna zdolność państwa musi rosnąć równolegle z proceduralną kontrolą sprawowaną przez obywatela. Audyt, prawo do informacji, kontrola sądowa, niezależne organy i polityczna odpowiedzialność stają się jeszcze ważniejsze, gdy administracja dysponuje bardziej kompetentną kognitywną infrastrukturą. AI może paradoksalnie zwiększyć znaczenie klasycznych zasad konstytucyjnych. Trójpodział władzy, jawność, kontrola sądowa i wolna debata są historycznymi technologiami przeciwdziałania nadmiernej koncentracji sprawczości. W epoce AI nie stają się anachronizmem, lecz mogą pełnić rolę epistemicznej redundancji społeczeństwa. Różne instytucje powinny posiadać odmienne informacje i kryteria oraz możliwość wzajemnego kwestionowania, aby nie powstał jeden system definiujący problem, dowód i decyzję.

Hybrydowa inteligencja jako świadomy problem projektowy

Ostatecznie pytanie o przyszłość AI prowadzi do teorii konstytucyjnej inteligencji zbiorowej. Najbardziej zaawansowane społeczeństwo nie będzie tym, które dysponuje najpotężniejszym pojedynczym modelem, lecz tym, które potrafi włączyć potężne systemy do struktury zachowującej różnorodność, prawa, sprawczość i zdolność samokorekty. Taka cywilizacja może osiągnąć znacznie wyższą inteligencję funkcjonalną niż dzisiejsza, bez konieczności przekształcania się w jeden superorganizm. W tym miejscu należy doprecyzować źródłowe pytanie o rolę człowieka jako instancji nadającej „sens i kierunek". Nie można bowiem z góry dowieść, że system sztuczny nigdy nie wykształci własnych znaczeń czy wartościowania.

Obecnie jednak normatywna legitymizacja decyzji społecznych spoczywa na ludziach i tworzonych przez nich instytucjach, gdyż to oni są podmiotami praw, interesów i politycznego mandatu. Ochrona ludzkiego kierunku nie jest więc metafizycznym dogmatem o wiecznej wyższości

człowieka, lecz aktualną zasadą polityczną: władza nad ludźmi musi być legitymizowana przez tych ludzi, których dotykają jej konsekwencje. Jeśli w przyszłości powstaną inne podmioty posiadające uzasadniony status moralny, architektura wspólnoty będzie wymagała rozszerzenia. Byłby to jednak kolejny problem konstytucyjny, a nie automatyczny rezultat benchmarku AGI. Tym samym kwestia praw świadomej AI nie powinna przesłaniać pilniejszego problemu: praw ludzi funkcjonujących już teraz w środowisku kształtowanym przez systemy pozbawione demokratycznego mandatu.

Właśnie tutaj odpowiedzialność dzisiejszych projektantów, ustawodawców, przedsiębiorców i obywateli nabiera wymiaru międzypokoleniowego. Infrastruktura przyjęta obecnie może stać się domyślnym środowiskiem dorastania ludzi, którzy nie zapamiętają świata sprzed generatywnej AI. Dla nich zewnętrzna pamięć, agentowa pomoc i syntetyczny partner poznawczy będą tak naturalne, jak wyszukiwarka czy smartfon dla wcześniejszych pokoleń. Odpowiedzialność polega zatem nie na próbie zachowania świata w obecnej formie, lecz na tym, jakie możliwości niezależności i krytyki zostawiamy ludziom, dla których hybrydowość będzie stanem zwyczajnym. Hybrydowa Ludzkość nie jest więc ani utopią, ani dystopią.

Jest problemem projektowym. Rezultat nie jest zdeterminowany technologicznie, lecz zależy od bieżących wyborów projektowych i instytucjonalnych. Teza ta stanowi kluczowy element inspiracji The AI Instinct, ponieważ odrzuca zarówno fatalizm technofilny, jak i technofobiczny. Nie musimy zakładać, że AI nas wyzwoli lub zniewoli.

Powinniśmy raczej badać mechanizmy, za pomocą których konkretna konfiguracja technologii, rynku, prawa i kultury wzmacnia jedne rodzaje sprawczości, a ogranicza inne. Zamiast pytać jedynie „jak potężna jest AI?”, należy pytać, kto dzięki niej staje się potężniejszy. Zamiast zastanawiać się, „czy AI jest inteligentna?” – jak zmienia ona strukturę inteligencji społeczeństwa. Zamiast pytać, „czy zastąpi pracowników?” – które kompetencje zostaną przeniesione z ludzi do infrastruktury i kto będzie jej właścicielem.

Zamiast rozważać, „czy będzie świadoma?” – jakie konsekwencje ma jej działanie jeszcze przed rozstrzygnięciem kwestii świadomości. Zamiast pytać, „czy osiągnie AGI?” – które funkcje poznawcze już dziś przekazujemy systemom i jak to zmienia przyszłych ludzi. Takie przeformułowanie nie umniejsza technicznych badań nad AGI; przeciwnie, osadza je w szerszym kontekście. Generalna kompetencja będzie niezwykle ważną cechą, jednak dla cywilizacji równie istotna okaże się jakość relacji między kompetencją a instytucją. Bardzo kompetentny system pozbawiony odpowiedzialności może być mniej wartościowy niż nieco słabszy model działający w dobrze zaprojektowanej hybrydzie. Najwyższa dokładność nie jest najważniejszym kryterium, jeśli system nie potrafi komunikować niepewności, poddawać się rewizji i współpracować z ludzkimi

procedurami. Fakt, że współczesne zespoły człowiek-AI nie osiągają automatycznie synergii, jest w tym kontekście korzystnym ostrzeżeniem: przyszłość hybrydowej inteligencji nie nastąpi sama. Trzeba ją zaprojektować.

W przeciwnym razie ryzykujemy stworzenie systemu jednocześnie bardziej skomplikowanego i mniej skutecznego, w którym człowiek zakłóca trafne decyzje AI, a AI destabilizuje intuicje człowieka. Najważniejszym przedmiotem przyszłych badań powinno być zatem nie tylko artificial intelligence, ale science of hybrid cognition. Dyscyplina ta musiałaby łączyć psychologię poznawczą z HCI, teorię decyzji z nauką o organizacji, neuronaukę z neuroetyką, uczenie maszynowe z teorią systemów, prawo z ekonomią polityczną oraz filozofię umysłu z epistemologią społeczną.

Jej jednostką analizy nie byłby model, lecz sprzężenie; podstawową zmienną nie tylko accuracy, lecz struktura przepływu informacji, kontroli i odpowiedzialności. Taka nauka wymagałaby nowych miar. Obok jakości wyniku należałoby mierzyć kalibrację zaufania, retained human capability, niezależność błędów, jakość kontradyktoryjności, przenośność pamięci, rozkład odpowiedzialności, zdolność systemu do eskalacji własnej niewiedzy oraz długoterminowy wpływ na kompetencje użytkownika. Na poziomie społecznym konieczny byłby pomiar koncentracji infrastruktury, różnorodności źródeł, stopnia kognitywnego lock-in, odporności na wspólne błędy i dostępności alternatywnych modeli. Dopiero wtedy „wydajność AI" przestanie być kategorią oderwaną od struktury społeczeństwa.

> Czy w pogoni za absolutną efektywnością nie budujemy systemu, który będzie perfekcyjnie trafny, lecz całkowicie pozbawiony zdolności do autorefleksji? Prawdziwym ryzykiem nie jest to, że maszyny staną się zbyt inteligentne, byśmy mogli nimi sterować, lecz że my staniemy się zbyt zależni od nich, by wiedzieć, jak myśleć samodzielnie. Ostatecznie przyszłość hybrydowej inteligencji nie jest przeznaczeniem, które nas spotka, lecz decyzją projektową, którą musimy podjąć teraz.

Inspiracje

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

Rana Gujral (ur. 18 czerwca 1976 r.) to indyjsko-amerykański przedsiębiorca technologiczny, inwestor i dyrektor biznesowy, specjalizujący się w kognitywnej sztucznej inteligencji i przetwarzaniu głosu. Ukończył informatykę i inżynierię na Uniwersytecie Mahatma Jyotiba Phule Rohilkhand oraz szkolenie dla kadry kierowniczej w MIT Sloan School of Management. Gujral jest najbardziej znany jako dyrektor generalny Behavioral Signals, firmy zajmującej się rozpoznawaniem emocji i analizą behawioralną, która wykorzystuje przetwarzanie sygnałów mowy do wnioskowania o intencjach, emocjach i ryzyku oszustwa na podstawie danych z ludzkiego głosu. Wcześniej założył firmę oprogramowania korporacyjnego TiZE (przejętą przez Alchemy) i zajmował stanowiska kierownicze w firmach Logitech i Cricut. Jako wybitny mówca branżowy i autor artykułów w publikacjach technologicznych i biznesowych, takich jak Forbes i TechCrunch, Gujral koncentruje się na hybrydowym poznaniu, inteligencji emocjonalnej w maszynach oraz społecznych implikacjach powstających systemów sztucznej inteligencji.

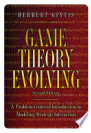
Rana Gujral *(born June 18, 1976) is an Indian-American technology entrepreneur, investor, and business executive specializing in cognitive artificial intelligence and voice computing. He holds a degree in computer science and engineering from Mahatma Jyotiba Phule Rohilkhand University and completed executive training at the MIT Sloan School of Management. Gujral is best known as the chief executive officer of Behavioral Signals, an emotion recognition and behavioral analytics company that applies speech-signal processing to infer intent, emotion, and deception risk from human voice data. Previously, he founded the enterprise software company TiZE (acquired by Alchemy) and held leadership positions at Logitech and Cricut. A prominent industry speaker and contributor to technology and business publications like Forbes and TechCrunch, Gujral focuses on hybrid cognition, emotional intelligence in machines, and the societal implications of emerging AI systems.*

Behavioral Signals

Literatura uzupełniająca

Książki

Książki autorów przywoływanych i literatura pokrewna:



[1] "game theory"

John Maynard Smith

Princeton University Press | ISBN: 9780691140513

[2] "The Theory of Evolution"

John Maynard Smith



[3] "Evolution and the Theory of Games"

John Maynard Smith

Cambridge University Press | ISBN: 9780521288842



[4] "The Sage Handbook of Human Resource Development"

Tonette S. Rocco, Michael Lane Morris, Rob F. Poell

2024 | SAGE Publications Limited | ISBN: 9781529678864

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Sustainable Human Development: Concepts and Priorities"

Sudhir Anand and Amartya Sen

1994 | RePEc: Research Papers in Economics

Źródło

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

Sudhir Anand and Amartya Sen

1995 | RePEc: Research Papers in Economics

Źródło

[3] "The Major Transitions in Evolution"

John Maynard Smith, Eörs Szathmáry

1997 | Oxford University Press eBooks | DOI: 10.1093/oso/9780198502944.001.0001

Źródło

[4] "The major evolutionary transitions"

Eörs Szathmáry, John Maynard Smith

1995 | Nature | DOI: 10.1038/374227a0

Źródło

[5] "The major transitions"

John Maynard Smith, EöRs SzathmáRy

1999 | The Origins of Life | DOI: 10.1093/oso/9780198504931.003.0002

Źródło

[6] "From Replicators to Reproducers: the First Major Transitions Leading to Life"

Eörs Szathmáry, John Maynard Smith

1997 | Journal of Theoretical Biology | DOI: 10.1006/jtbi.1996.0389

Źródło

[7] "The Major Transitions of Evolution."

Egbert Giles Leigh, John Maynard Smith, Eörs Szathmáry

1995 | Evolution | DOI: 10.2307/2410462

Źródło

[8] "Toward major evolutionary transitions theory 2.0"

Eörs Szathmáry

2015 | Proceedings of the National Academy of Sciences | DOI: 10.1073/pnas.1421398112

Źródło

Artykuły naukowe autora:

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch02

Źródło

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch03

Źródło

[11] "The AI Instinct"

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.fmatter

Źródło

[12] **"The Emergence of Artificial General Intelligence (AGI)"**

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch04

Źródło

[13] **"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: 370fa646-2cbd-471f-a44e-54a8cb64f037