Biologiczna architektura kognicji a instynkt AI w świetle koncepcji Rany Gujral

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Tekst analizuje ludzką kognicję nie jako wadliwy komputer, lecz jako zoptymalizowany biologicznie system regulujący relację między organizmem a środowiskiem. Opierając się na teorii ograniczonej racjonalności Herberta Simona oraz badaniach Kahnemana i Tversky'ego, autor argumentuje, że ludzkie heurystyki nie są błędami logicznymi, lecz efektywnymi adaptacjami do świata o skończonych zasobach. W przeciwieństwie do AI człowiek nie dąży do maksymalizacji użyteczności, lecz do satysfakcji (satisficing), co wynika z metabolicznych i czasowych kosztów przetwarzania informacji. Różnica między systemem biologicznym a sztucznym leży zatem w odmiennej architekturze kosztów i historii ewolucyjnej, a nie w samym fakcie bycia ograniczonym.

The text analyzes human cognition not as a flawed computer, but as a biologically optimized system regulating the relationship between the organism and its environment. Drawing on Herbert Simon's theory of bounded rationality and the research of Kahneman and Tversky, the author argues that human heuristics are not logical errors, but efficient adaptations to a world of finite resources. Unlike AI, humans do not strive for utility maximization, but for 'satisficing,' which results from the metabolic and temporal costs of information processing. The difference between biological and artificial systems thus lies in different cost architectures and evolutionary histories, rather than in the mere fact of being limited.

Artykuł stanowi krytyczną analizę ludzkiej kognicji, odrzucając redukcjonistyczne wizje człowieka jako „wadliwego komputera” lub nośnika tajemniczej intuicji. Autor stawia tezę, że biologiczne ograniczenia – takie jak selektywność uwagi, emocje czy

ucieleśnienie – nie są błędami systemu, lecz fundamentalnymi elementami architektury adaptacyjnej, która umożliwia sprawne funkcjonowanie w nieprzewidywalnym świecie. Tekst analizuje mechanizmy ograniczonej racjonalności i metapoznania, wskazując na ryzyko „dryfu sprawczości" w relacji z AI, gdzie wygoda outsourcingu poznawczego może prowadzić do utraty zdolności krytycznego nadzoru. Kluczowym punktem rozważań jest różnica między kompetencją a mądrością; podczas gdy AI może dorównać człowiekowi w szerokości umiejętności, brakuje jej ucieleśnionej historii i biologicznej osi doświadczenia, które dla ludzi są fundamentem nadawania znaczeń. W konsekwencji artykuł prowokuje do pytania o to, czy inteligencja bez biologicznego zakotwiczenia i potrzeby przetrwania może kiedykolwiek wykształcić prawdziwy osąd.

SŁOWA KLUCZOWE

biologiczna architektura kognicji | ograniczona racjonalność | metapoznanie | cognitive offloading | deskilling | epistemiczne tarcie | dryf sprawczości | heurystyki poznawcze | neuroplastyczność | systemy hybrydowe człowiek-AI | model AGE | architektura nagrody | teoria perspektywy | funkcjonalny odpowiednik doświadczenia | kalibracja zaufania

Kognicja jako racjonalność osadzona w biologicznych ograniczeniach

Biologiczna architektura ludzkiej kognicji: ograniczona racjonalność, ciało, emocje, pamięć i metapoznanie

Skoro granica umysłu została rozszerzona poza prostą lokalizację procesów poznawczych w mózgu, kolejnym krokiem musi być precyzyjne określenie tego, co właściwie wnosi do procesu poznawczego sam człowiek. Nie wystarczy stwierdzić, że jest on „intuicyjny", „emocjonalny" czy „ucieleśniony". Bez rekonstrukcji konkretnych mechanizmów takie sformułowania łatwo zamieniają analizę naukową w antropologiczny romantyzm, wedle którego biologiczna niedoskonałość człowieka miałaby automatycznie stanowić źródło głębszej mądrości, niedostępnej maszynie. Równie nietrafny obraz wyłania się z przeciwnego bieguna, gdy człowieka przedstawia

się jako wadliwy komputer: system obciążony biasami, zbyt małą pamięcią roboczą, emocjonalnymi zakłóceniami i niewystarczającą szybkością obliczeń. Obie te metafory są redukcjonistyczne.

Biologiczny system poznawczy nie jest ani niedoskonałą wersją algorytmu optymalizacyjnego, ani tajemniczym nośnikiem irracjonalnej intuicji. Jest organizmem, którego mechanizmy poznawcze wyewoluowały jako element systemu regulującego relację między stanem wewnętrznym, środowiskiem a działaniem. Z tego powodu ograniczenia, emocje, selektywność uwagi, heurystyki i niepewność nie są jedynie zakłóceniami „czystego rozumu", lecz integralnymi częściami architektury umożliwiającej funkcjonowanie w świecie, którego nie da się w całości przeliczyć przed podjęciem decyzji. Punkt wyjścia stanowi tutaj teoria ograniczonej racjonalności Herberta Simona. Jej znaczenie polegało nie tyle na stwierdzeniu oczywistego dziś faktu, że człowiek popełnia błędy, co na zakwestionowaniu modelu decydenta jako podmiotu dysponującego kompletną informacją, stabilnym zbiorem preferencji oraz mocą obliczeniową pozwalającą wybrać rozwiązanie maksymalizujące użyteczność. Człowiek operuje pod presją czasu, kosztów pozyskania danych oraz ograniczeń uwagi i pamięci. W praktyce często nie maksymalizuje, lecz satisficuje: poszukuje rozwiązania wystarczająco dobrego w stosunku do aspiracji i warunków sytuacyjnych.

Ograniczona racjonalność nie jest zatem teorią irracjonalności, lecz teorią racjonalności osadzonej w realnych warunkach funkcjonowania systemu o skończonych zasobach. W kontekście porównania człowieka z AI jest to kluczowe, ponieważ system sztuczny również nigdy nie operuje na „całej rzeczywistości"; działa w ramach określonych danych, kontekstu i budżetu obliczeniowego, z konkretną funkcją celu i ograniczoną zdolnością przewidywania konsekwencji. Różnica między człowiekiem a maszyną nie sprowadza się zatem do przeciwstawienia ograniczonego podmiotu biologicznego nieograniczonemu podmiotowi obliczeniowemu. Oba systemy są ograniczone, lecz przez inne zasoby, inną historię powstania i odmienną architekturę kosztów.

Psychologia sądu i decyzji rozwinęła ten problem w badaniach nad heurystykami. Prace Amosa Tversky'ego i Daniela Kahnemana wykazały, że ludzkie szacowanie prawdopodobieństwa, ryzyka i wartości często odbiega od normatywnych modeli teorii prawdopodobieństwa oraz oczekiwanej użyteczności. Heurystyka dostępności może zawyżać ocenę zdarzeń łatwo przywoływanych z pamięci; reprezentatywność skłania do pomijania częstości bazowych, a zakotwiczenie sprawia, że początkowa wartość wpływa na późniejsze oszacowania, nawet gdy nie powinna mieć znaczenia. Teoria perspektywy pokazała ponadto, że wybory ludzi zależą nie tylko od końcowych stanów majątkowych, ale od punktów odniesienia, subiektywnej wartości zysków i strat oraz asymetrii między nimi.

Heurystyki jako efektywna adaptacja biologiczna

W kulturze popularnej wyniki tych badań bywają streszczane stwierdzeniem, że „człowiek jest irracjonalny". Jest to jednak wniosek zbyt szeroki. Badania wykazały przede wszystkim systematyczne odstępstwa od norm formalnych w konkretnych warunkach eksperymentalnych; nie dowodzą one, że ludzkie poznanie jako całość jest źle zaprojektowaną procedurą rachunkową.

Współczesna perspektywa ekologiczna pozwala pójść dalej. Heurystyka może prowadzić do błędu w odniesieniu do abstrakcyjnego standardu, a jednocześnie działać efektywnie w środowisku, którego statystyczna struktura odpowiada warunkom jej powstania. Oznacza to, że funkcję mechanizmu poznawczego należy oceniać nie tylko przez pryzmat logicznej poprawności pojedynczego sądu, lecz także poprzez relację między kosztem obliczeń, dostępnością danych, czasem reakcji a konsekwencjami pomyłki. Mózg jest narządem kosztownym metabolicznie, selektywnie kodującym i przetwarzającym informacje. Modele łączące ekonomię informacji, neurobiologię i teorię efektywnego kodowania wskazują, że percepcję oraz preferencje trzeba analizować z uwzględnieniem ograniczeń biologicznych i stanów wewnętrznych organizmu; system sensoryczny nie jest neutralnym urządzeniem, którego celem jest maksymalnie wierne odtworzenie całej dostępnej informacji.

Z tej perspektywy „skrót poznawczy" nie musi być wadą. Może być ceną, jaką system płaci za możliwość działania przed uzyskaniem kompletnej wiedzy. Rozróżnienie Systemu 1 i Systemu 2 stanowi tutaj użyteczny język analizy kognicji: pierwszy tryb jest szybki, automatyczny i intuicyjny, drugi zaś wolniejszy, wymagający uwagi i kontroli. Takie ujęcie łączy kwestię ograniczonej racjonalności i heurystyk z późniejszym porównaniem człowieka z AI. Należy jednak zachować precyzję, której często brakuje w popularyzacji teorii dual-process. System 1 i System 2 nie są dwoma organami psychicznymi ani odrębnymi anatomicznie modułami. Współczesna literatura przedmiotu obejmuje konkurujące modele, a sama deliberacja okazuje się zjawiskiem wewnętrznie zróżnicowanym.

Wim De Neys proponuje rozróżnienie co najmniej czterech funkcji deliberacji: kontroli odpowiedzi, generowania alternatywnej odpowiedzi, uzasadniania oraz regulacji przebiegu rozumowania. Przeciwstawienie szybkiego instynktu wolnej refleksji jest więc użyteczną mapą, lecz nie pełną teorią architektury umysłu. Co istotne, proces deliberatywny nie zawsze kończy się gwałtownym przełączeniem z niepewności do decyzji. Badania neurofizjologiczne wskazują, że reprezentacja dowodów leżących u podstaw wyboru może zachowywać charakter stopniowany aż do momentu podjęcia decyzji. W badaniu z zapisami wewnątrzczaszkowymi opublikowanym w 2024 roku wykazano stopniowe narastanie sygnału związanego z gromadzeniem dowodów, co sugeruje, że

decyzja nie musi być wyłącznie binarnym przekroczeniem stałego progu. Choć na podstawie jednego eksperymentu nie buduje się uniwersalnych teorii, wynik ten dobrze ilustruje ważniejszą zasadę: decyzja, która w sferze zachowania przybiera postać „tak" lub „nie", może wyrastać z ciągłego rozkładu niepewności, konkurujących reprezentacji i zmiennej siły dowodów. Świadomy wybór jest końcowym punktem procesu znacznie bogatszego niż jego publicznie obserwowalny rezultat. To prowadzi nas do problemu uwagi.

Organizm otrzymuje więcej informacji, niż jest w stanie przetworzyć z jednakową dokładnością. Uwaga nie jest więc wyłącznie zdolnością do „koncentracji", lecz mechanizmem alokowania ograniczonych zasobów poznawczych. Selekcja określa, które bodźce otrzymają dostęp do dalszego przetwarzania, które cele zyskają pierwszeństwo i które elementy sytuacji zostaną utrzymane wystarczająco długo, by wpłynąć na zachowanie. Już na tym poziomie metafora racjonalnego decydenta okazuje się niewystarczająca. Decyzję można podjąć wyłącznie na podstawie reprezentacji, które przedostały się do procesu decyzyjnego. To, czego system nie zauważył, czego nie utrzymał w pamięci i czemu nie przypisał znaczenia, nie będzie równorzędnym kandydatem w późniejszej deliberacji. W konsekwencji władza nad uwagą jest pośrednio władzą nad zbiorem przesłanek, z którego powstaje osąd.

Doświadczenie jako strukturalna transformacja systemu

Ta ludzka właściwość nabiera szczególnego znaczenia w epoce systemów rekomendacyjnych i generatywnej AI. Skoro dostęp do uwagi stanowi wąskie gardło ludzkiej kognicji, system nie musi „przekonywać" użytkownika po zakończeniu racjonalnej analizy, by wpłynąć na jego zachowanie. Może on wcześniej zmienić rozkład informacji poddawanych tej analizie. W tym punkcie psychologia poznawcza łączy się z ekonomią uwagi i teorią infrastrukturalnej władzy algorytmicznej. Mechanizm ten opisuje przejście AI od kształtowania outputs do kształtowania inputs umysłu: tego, co widzimy, czytamy, uznajemy za istotne i włączamy do procesu rozumowania. Nie trzeba zatem przyjmować radykalnej tezy o bezpośrednim „programowaniu człowieka", by dostrzec realny mechanizm wpływu. Zmiana środowiska informacyjnego modyfikuje zachowanie, ponieważ ludzka racjonalność operuje zawsze na materiale wcześniej wyselekcjonowanym.

Drugim kluczowym mechanizmem selekcyjnym jest pamięć. Powszechna metafora traktuje ją jak magazyn, z którego wiernie pobiera się dane. Współczesna psychologia i neuronauka widzą ją jednak jako proces dynamiczny i rekonstrukcyjny. Uczenie zmienia system, a przywołanie wspomnienia nie jest prostym odczytem niezmiennego zapisu; informacja podlega organizacji,

integracji z wcześniejszą wiedzą, zapominaniu, generalizacji i rekonstrukcji. Z ewolucyjnego punktu widzenia taki system nie musiał być bezbłędnym archiwum przeszłości. Jego wartość tkwi w umożliwianiu przewidywania, planowania i dostosowywania przyszłych zachowań. Wspomnienie jest użyteczne nie jako kopia historii, lecz jako narzędzie zmieniające przyszłą reakcję.

Tutaj pojęcie doświadczenia nabiera precyzyjniejszego znaczenia. Doświadczenie nie jest sumą zgromadzonych rekordów. Osoba doświadczona nie dysponuje jedynie większą bazą danych; w toku interakcji zmieniła sposób rozpoznawania sytuacji, hierarchię istotności sygnałów, oczekiwania, repertuar strategii i ocenę ryzyka. Doświadczenie jest historią, która stała się strukturą bieżącego systemu. W książce ta intuicja służy do zbudowania modelu AGE: pamięć pojmowana jest tu nie jako storage, lecz jako transformacja, dzięki której system po konsekwencjach własnego działania przestaje być tym samym systemem. Teza ta wymaga późniejszego rozbicia na poziomy uczenia parametrów, zmiany polityki działania, pamięci autobiograficznej, modelu siebie, utrzymania tożsamości oraz ewentualnego doświadczenia fenomenalnego. Na poziomie biologicznej kognicji fundament tej tezy jest jednak jasny: przeszłość wpływa na przyszłość poprzez zmianę układu, który poznaje.

Tę transformację nie da się zrozumieć bez neuroplastyczności. Termin ten obejmuje zdolność układu nerwowego do zmian strukturalnych i funkcjonalnych wynikających z rozwoju, uczenia, adaptacji, uszkodzeń i doświadczenia. Plastyczność nie oznacza jednak nieograniczonej dowolności przebudowy mózgu – zależy ona od wieku, regionu, rodzaju bodźców, stanu biologicznego i historii jednostki.

Ryzyko deskilling i utrata kontroli nad delegowanymi kompetencjami

Plastyczność nie jest również normatywnie „dobra” sama w sobie: mózg może utrwalać korzystne umiejętności, ale równie dobrze maladaptacyjne nawyki. Technologia, wpływając na częstotliwość wykonywania określonych czynności, schemat uwagi i rodzaj wysiłku poznawczego, z czasem zmienia zatem kompetencje użytkownika. Jednocześnie dowody dotyczące wpływu współczesnych narzędzi cyfrowych na funkcje poznawcze są bardziej złożone, niż sugeruje katastroficzna publicystyka. Przegląd z 2024 roku wskazuje zarówno na możliwe korzyści, jak i zagrożenia, podkreślając przy tym, że wiedza o dokładnych mechanizmach neurobiologicznych pozostaje ograniczona. Nie ma więc podstaw do automatycznego twierdzenia, że korzystanie z AI „niszczy mózg”.

Istnieją natomiast powody, by badać, które funkcje są ćwiczone, a które delegowane i jakie długoterminowe skutki przynosi dany wzorzec współpracy. Problem ten można ująć w ramach teorii cognitive offloading. Człowiek od dawna przenosi część obciążeń poznawczych na środowisko: zapisuje informacje zamiast je zapamiętywać, używa kalkulatora zamiast rachunków w pamięci, korzysta z mapy zamiast budować precyzyjny wewnętrzny model przestrzeni. Delegowanie nie jest samo w sobie degeneracją poznania; może wręcz uwolnić zasoby niezbędne do zadań wyższego rzędu.

Matematyk używający oprogramowania nie musi stawać się gorszym matematykiem, podobnie jak naukowiec korzystający z bibliografii elektronicznej nie traci zdolności argumentowania. Problem pojawia się wtedy, gdy delegujemy nie tylko czynność pomocniczą, lecz także kompetencję niezbędną do oceny rezultatu. Jeśli użytkownik przestaje rozumieć proces na tyle, by rozpoznać błąd systemu, powstaje asymetryczna zależność: narzędzie podnosi krótkoterminową wydajność, ale jednocześnie osłabia zdolność do niezależnej kontroli. Właśnie dlatego pojęcie deskilling jest tu istotniejsze niż ogólne hasło „atrofii". W historii technologii wielokrotnie dochodziło do zaniku pewnych umiejętności wraz ze zmianą podziału pracy. Nie oznaczało to zazwyczaj globalnego spadku inteligencji, lecz zmianę repertuaru kompetencji potrzebnych jednostce. Automatyzacja nawigacji zmniejsza zapotrzebowanie na pewne operacje przestrzenne, autokorekta redukuje konieczność pamiętania reguł pisowni, a generatywna AI może ograniczyć częstotliwość samodzielnego tworzenia pierwszych wersji tekstu, kodu czy analizy.

Zasadnicze pytanie brzmi: czy utracona kompetencja jest zastępowana przez zdolność wyższego rzędu – na przykład umiejętność oceny, syntezy i nadzoru – czy też wraz z wykonaniem zadania delegujemy również kryterium pozwalające rozpoznać jego poprawność? W drugim przypadku człowiek może zachowywać formalne prawo do decyzji przy malejącej materialnej zdolności do jej kontrolowania.

Metapoznanie jako klucz do sprawczości w relacji z AI

Z neuroplastycznością wiąże się kolejny problem: architektura nagrody. Organizmy uczą się nie tylko poprzez jawne przyswajanie reguł, ale także przez wzmacnianie działań przynoszących określone skutki. Układ dopaminergiczny to coś więcej niż prosty „system przyjemności"; odpowiada on m.in. za uczenie sygnałów predykcyjnych, motywację i aktualizację oczekiwań względem nagrody. Popularny język "dopaminowych strzałów" upraszcza tę złożoną neurobiologię, jednak dla analizy środowiska cyfrowego wystarczy teza: częstotliwość, przewidywalność i czasowa bliskość informacji zwrotnej kształtują wzorce uczenia się i zachowania. System, który stale skraca

dystans między potrzebą a odpowiedzią, może trwale zmienić oczekiwania użytkownika co do kosztu poznawczego. Nie trzeba zatem mówić o fizjologicznym „uzależnieniu od AI", by dostrzec ryzyko wykształcenia nawyku wybierania "ścieżki o najmniejszym wysiłku".

Kwestia kosztu poznawczego była centralna już dla klasycznych teorii dual-process. Deliberacja wymaga czasu, zasobów i kontroli, a uruchamianie najdroższego trybu analizy przy każdej decyzji byłoby nieefektywne. Pojawia się zatem pytanie metapoznawcze: kiedy warto myśleć dłużej? W praktyce skuteczność systemu zależy nie tylko od posiadanych zdolności analitycznych, lecz przede wszystkim od umiejętności rozpoznania momentu, w którym należy ich użyć. Ten problem nabiera szczególnego znaczenia w relacji człowiek-AI.

Jeśli system generuje płynną i pozornie kompletną odpowiedź, użytkownik otrzymuje sygnał, że dalsza deliberacja jest zbędna. Nie chodzi więc wyłącznie o poprawność odpowiedzi AI, lecz o to, jak sposób komunikowania pewności wpływa na decyzję człowieka o uruchomieniu własnej kontroli poznawczej. Prowadzi to bezpośrednio do metapoznania. Najprościej można je zdefiniować jako poznanie dotyczące własnych procesów poznawczych, choć pojęcie to obejmuje szeroki wachlarz zjawisk: wiedzę o własnych zdolnościach, monitorowanie wykonania, ocenę pewności, rozpoznawanie błędów, decyzję o dalszym poszukiwaniu informacji oraz regulację strategii. Współczesne badania rozróżniają subpersonalne reprezentacje niepewności od osobowego sądu w rodzaju: „jestem w 70 procentach pewna tej decyzji". Stephen Fleming podkreśla, że poczucie pewności nie jest prostym odczytem obiektywnej poprawności, lecz wynikiem wnioskowania o własnej decyzji na podstawie modelu świata i modelu własnego systemu poznawczego. W efekcie metapoznawcza pewność może być dobrze lub źle skalibrowana względem rzeczywistej skuteczności.

Właśnie ta właściwość odróżnia dwa poziomy kompetencji, które w potocznym języku często zlewają się w jedno. Można regularnie odpowiadać poprawnie, a jednocześnie błędnie szacować własną wiarygodność; można też dysponować przeciętną kompetencją pierwszego rzędu, lecz precyzyjnie wiedzieć, kiedy własna odpowiedź jest niepewna. W współpracy człowiek-AI jakość metapoznania może okazać się ważniejsza niż surowa liczba poprawnych odpowiedzi. Użytkownik nie musi przewyższać modelu w każdym zadaniu, by zachować sprawczość – musi natomiast umieć rozpoznać, kiedy mu zaufać, kiedy zażądać dowodu, kiedy zasięgnąć drugiej opinii, a kiedy odrzucić rekomendację. Analogiczny obowiązek spoczywa na systemie sztucznym: model użyteczny w decyzjach o wysokiej stawce powinien nie tylko generować rezultat, ale możliwie precyzyjnie reprezentować i komunikować własną niepewność. To właśnie na "styku dwóch niedoskonałych metapoznań" – ludzkiego i maszynowego – rodzi się jeden z kluczowych problemów bezpieczeństwa systemów hybrydowych.

Emocje i metapoznanie jako mechanizmy selekcji danych

Metapoznania nie powinno się jednak redukować do prostego wskaźnika confidence. Przegląd badań z 2024 roku wskazuje, że choć precyzyjne pomiary pewności były ogromnym osiągnięciem tej dziedziny, mogą one przesłonić szersze formy metawiedzy: długoterminowe przekonania o własnych kompetencjach, relację metapoznania z poznaniem społecznym czy sposób oceny stanów afektywnych. Dla naszej analizy jest to kluczowe. Hybrydowa kognicja nie polega bowiem na serii odizolowanych pytań egzaminacyjnych, w których użytkownik i model raportują procentową pewność, lecz jest relacją rozwijaną w czasie.

Człowiek może nauczyć się, że "AI zwykle wie lepiej", i przekształcić to doświadczenie w trwałą regułę delegowania – nawet tam, gdzie model nie dysponuje wystarczającymi danymi. Może też po kilku spektakularnych błędach przyjąć odwrotną, równie uproszczoną regułę nieufności. Metapoznanie jest zatem również historią kalibracji zaufania. Do tej samej architektury należy emocja. Przez znaczną część zachodniej historii idei rozum i emocja były przeciwstawiane: rozum miał kontrolować namiętności, które zakłócały trafny osąd. Współczesna psychologia decyzji znacznie komplikuje ten obraz. Emocje wpływają na uwagę, pamięć, ocenę prawdopodobieństwa, postrzeganie ryzyka, motywację i wybór, nie działając przy tym jak jeden ogólny suwak "racjonalności".

Badania Jennifer Lerner i współpracowników pokazują, że różne emocje wywierają odmienne, częściowo przewidywalne efekty, zależne od struktur oceny sytuacji, takich jak poczucie kontroli i pewności. Emocja może pogarszać wynik w jednym kontekście i poprawiać adaptację w innym; wpływ afektu na decyzje jest powszechny, ale nie jednokierunkowy. Gujral odwołuje się do tej tradycji, przypominając m.in. różnice między wpływem lęku, złości i smutku na przetwarzanie ryzyka i informacji. Szczególnie istotne jest rozróżnienie emocji integralnych i incydentalnych. Emocja integralna wynika z samego przedmiotu decyzji: lęk przed niebezpieczną inwestycją lub zabiegiem medycznym może nieść informację o postrzeganym ryzyku. Emocja incydentalna pochodzi z innego źródła, lecz może zostać błędnie przeniesiona na bieżący sąd. Ta sama zdolność afektywna może więc wspomagać wartościowanie albo prowadzić do błędu atrybucji. Na poziomie teorii kognicji oznacza to, że emocji nie można po prostu "usunąć", aby odsłonić czyste dane. System decyzyjny potrzebuje mechanizmu określającego hierarchię ważności danych; w biologicznym organizmie afekt jest jednym z narzędzi tej selekcji.

Próba skonstruowania "całkowicie racjonalnej" inteligencji poprzez pozbawienie jej funkcjonalnego odpowiednika wartościowania może więc prowadzić nie do doskonałego decydenta, lecz do

systemu posiadającego ogromną przestrzeń możliwych obliczeń bez wystarczającego mechanizmu wyboru, którym z nich nadać pierwszeństwo.

Hipoteza markerów somatycznych Antonia Damasio była jedną z najbardziej wpływowych prób wyjaśnienia związku pomiędzy stanem ciała, emocją i wyborem. Zgodnie z jej zasadniczą intuicją reprezentacje stanów cielesnych związanych z wcześniejszymi konsekwencjami mogą wpływać na ocenę dostępnych możliwości, ograniczając przestrzeń decyzji zanim każda alternatywa zostanie w pełni świadomie przeliczona. Klasyczne obserwacje pacjentów z uszkodzeniami obszarów przyśrodkowej kory przedczołowej wskazywały, że zachowana zdolność logicznego rozumowania nie gwarantuje skutecznego podejmowania decyzji w sytuacjach wymagających wartościowania. Gujral wykorzystuje ten model jako argument przeciw traktowaniu emocji jako "szumu" oraz jako podstawę tezy o znaczeniu biologicznego wartościowania dla inteligencji. Trzeba jednak zachować metodologiczną ostrożność.

Silna wersja hipotezy markerów somatycznych była przedmiotem poważnej krytyki, szczególnie w związku z interpretacją Iowa Gambling Task, możliwością świadomego rozpoznawania reguł zadania oraz niejednoznacznością dowodów na przyczynową rolę obwodowych sygnałów fizjologicznych. Krytyczny przegląd literatury wskazywał, że część obserwacji można wyjaśniać konkurencyjnymi mechanizmami i że teoria wymaga bardziej precyzyjnych dowodów. Nie podważa to szerszej, znacznie lepiej ugruntowanej tezy, że procesy emocjonalne i cielesne wpływają na decyzje. Uniemożliwia natomiast przedstawianie jednej konkretnej hipotezy jako zakończonego dowodu, iż "ciało podejmuje decyzję przed umysłem".

Biologiczna homeostaza jako fundament realnego znaczenia

Szerszym i obecnie bardzo produktywnym polem badań jest interocepcja, czyli przetwarzanie sygnałów dotyczących wewnętrznego stanu organizmu. Obejmuje ona między innymi informacje o układzie krążenia, oddychaniu, temperaturze, metabolizmie, potrzebach fizjologicznych oraz sygnałach trzewnych. Interocepcja uczestniczy w homeostazie, lecz jej znaczenie wykracza poza czysto fizjologiczną regulację: stan ciała współkształtuje emocje, motywacje i decyzje. Współczesne badania podkreślają jednocześnie, że mechanizmy integracji sygnałów interoceptywnych z informacją zewnętrzną wciąż pozostają przedmiotem intensywnych analiz. Dla teorii AI jest to istotne nie dlatego, że każda inteligencja musi posiadać żołądek, serce i układ hormonalny, lecz ponieważ biologiczna inteligencja dysponuje stale aktualizowanym wektorem stanu własnego

organizmu, który nadaje kierunek jego zachowaniu. Homeostaza dostarcza tu podstawowego wzorca.

Żywy organizm nie jest systemem rozwiązującym dowolne problemy na zlecenie zewnętrznego operatora. Jego istnienie zależy od utrzymania parametrów w zakresie umożliwiającym dalsze funkcjonowanie. Potrzeba tlenu, odpowiedniej temperatury, energii, bezpieczeństwa i integralności fizycznej tworzy naturalną asymetrię pomiędzy stanami świata. Niektóre są dla systemu obojętne, inne korzystne, a jeszcze inne zagrażają jego ciągłości. Zanim pojawi się świadoma teoria wartości, biologia już wartościuje w minimalnym sensie regulacyjnym. To właśnie ta zależność między stanem świata a możliwością trwania organizmu sprawia, że pojęcia „kosztu" i „konsekwencji" mają dla człowieka treść, której nie można utożsamiać z arbitralnie nadaną wartością funkcji celu. Nie wynika z tego jednak, że sztuczny system nigdy nie mógłby posiadać funkcjonalnego odpowiednika homeostazy.

Robot może monitorować stan baterii, temperaturę podzespołów, integralność sensorów czy prawdopodobieństwo awarii, a agent programowy zarządzać własnym budżetem obliczeniowym. Można zatem skonstruować system, w którym określone stany techniczne będą warunkiem dalszego działania. Różnica polega na tym, że w obecnych systemach cele te są projektowane i instrumentalne, podczas gdy w organizmie biologicznym regulacja własnej ciągłości jest wpisana w ewolucyjną historię systemu. Czy różnica genezy jest istotna dla inteligencji, autonomii i doświadczenia, pozostaje kwestią otwartą.

Nie należy jej rozstrzygać przez definicję. Pytanie to stanie się później jednym z najważniejszych problemów AGE: czy wystarczy stworzyć funkcjonalny odpowiednik „tego, co dla systemu ma znaczenie", czy doświadczenie wymaga również subiektywnego przeżywania tego znaczenia? Do obrazu ludzkiej kognicji należy dodać stres. Reakcja stresowa nie jest prostym patologicznym zakłóceniem; w krótkim okresie potrafi reorganizować uwagę, mobilizować zasoby energetyczne i zwiększać gotowość do działania w obliczu zagrożenia. Przewlekła lub skrajna aktywacja tych mechanizmów może natomiast wpływać na pamięć, funkcje wykonawcze, zdrowie somatyczne i regulację emocjonalną. Znaczenie stresu dla naszej analizy polega na tym, że ponownie ukazuje on poznanie jako funkcję organizmu posiadającego historię. Ta sama sytuacja może być oceniona odmiennie w zależności od wcześniejszych doświadczeń, poczucia kontroli, stanu fizjologicznego i wyuczonych oczekiwań. Racjonalność biologiczna nie rozpoczyna każdej decyzji od stanu zerowego.

Gujral idzie dalej i przywołuje badania nad potomkami osób, które przeżyły Holocaust, interpretując je jako przykład epigenetycznego przenoszenia fizjologicznych konsekwencji traumy między pokoleniami. W tym miejscu konieczne jest wyraźne ograniczenie tezy.

Badania na ludziach wykazały związki między traumą rodziców a różnicami w markerach epigenetycznych u potomstwa – między innymi w genach odpowiedzialnych za regulację osi stresowej – lecz nie stanowią one dowodu prostego mechanizmu dziedziczenia „pamięci traumy". Przegląd literatury wskazuje, że wyniki są wstępne i częściowo niespójne, a interpretację utrudniają małe próby, różnice metodologiczne, środowisko prenatalne, zachowania rodzicielskie oraz inne kanały transmisji społecznej. Naukowo bezpieczniejsze jest zatem stwierdzenie, że historia poprzednich pokoleń może wpływać na rozwój jednostki poprzez splot mechanizmów społecznych, psychologicznych, prenatalnych i biologicznych, podczas gdy dokładna rola dziedziczenia epigenetycznego u człowieka pozostaje przedmiotem badań. Ta korekta nie osłabia zasadniczej argumentacji.

Wielowarstwowa historia doświadczenia jako fundament ludzkiej inteligencji

Przeciwnie, widać tu, że ludzka kognicja posiada historię rozgrywającą się na wielu poziomach jednocześnie. Jednostka otrzymuje genom, rozwija się w określonym środowisku prenatalnym, jest socjalizowana przez osoby z własnymi doświadczeniami, wchodzi w struktury instytucji i kultury o długiej pamięci zbiorowej oraz przyswaja język przechowujący historyczne kategorie, by następnie samej podlegać uczeniu. Tożsamość poznawcza nie jest więc wynikiem pojedynczego strumienia danych, lecz przecięciem historii biologicznej, rozwojowej, rodzinnej, społecznej i kulturowej. Właśnie dlatego proste zestawienie "liczby parametrów AI" z "liczbą neuronów człowieka" niewiele mówi o rzeczywistej różnicy między tymi systemami.

Ich stan początkowy oraz mechanizmy aktualizacji wynikają z zupełnie innych procesów historycznych. Kolejnym elementem tej historii jest rozwój, w którym ludzkie kompetencje nie pojawiają się jednocześnie. Układ nerwowy dojrzewa w kontakcie z ciałem, opiekunami, językiem, przestrzenią społeczną i materialnym otoczeniem. Dziecko zdobywa wiedzę o świecie poprzez tysiące powtarzających się pętli percepcja-działanie-konsekwencja. Uczy się nie tylko kategorii przedmiotów, lecz także ich affordances: tego, co można z nimi zrobić. Równocześnie uczy się przewidywać zachowania innych osób, dostosowywać emocje, rozpoznawać normy i negocjować własne cele. Rozwój poznawczy jest zatem procesem jednoczesnego nabywania kompetencji sensorycznych, ruchowych, językowych, emocjonalnych, społecznych i normatywnych.

Oddzielenie tych warstw bywa użyteczne analitycznie, lecz nie oznacza ich niezależności w realnym rozwoju. Ta obserwacja jest kluczowa dla porównania z uczeniem maszynowym. Współczesny model językowy może przetwarzać ilość tekstu wielokrotnie przekraczającą to, co człowiek

przeczyta w ciągu życia, jednak relacja modelu z tym materiałem różni się od relacji dziecka z doświadczeniem. Człowiek uczy się znaczenia słowa "gorące" w sieci powiązań językowych, cielesnych, społecznych i pragmatycznych; model natomiast przyswaja wysoce precyzyjne zależności użycia tego słowa w ogromnym korpusie danych.

Oba systemy mogą później poprawnie posługiwać się tym terminem, lecz droga prowadząca do tej kompetencji jest odmienna. Z tego powodu nie wolno automatycznie uznać biegłości językowej modelu za dowód tożsamego rozumienia, ale nie należy też z góry odrzucać możliwości, że relacyjne struktury wyuczone z języka mogą realizować funkcje wcześniej przypisywane wyłącznie doświadczeniu sensorycznemu. Równie istotne jest poznanie społeczne. Człowiek nie tylko przewiduje fizyczny świat, lecz żyje w środowisku złożonym z innych podmiotów posiadających zamiary, przekonania, emocje i normatywne oczekiwania. Znaczna część ludzkiej inteligencji służy koordynowaniu współpracy, rozpoznawaniu wiarygodności, negocjowaniu statusu, wykrywaniu oszustwa, uczeniu się przez obserwację i przewidywaniu reakcji innych osób. To środowisko sprawia, że wiele informacji ma charakter społecznie pochodny. Wiemy ogromną część tego, czego sami nigdy nie sprawdziliśmy, ponieważ ufamy świadectwu ekspertów, instytucjom, podręcznikom i procedurom wytwarzania wiedzy. Poznanie jednostki od dawna jest więc epistemicznie zależne od innych ludzi.

AI wchodzi w to miejsce nie jako pierwszy zewnętrzny autorytet, lecz jako nowy rodzaj pośrednika wiedzy. Różnica polega na skali, szybkości i formie syntetyzowania informacji. Człowiek przyzwyczajony do ufania społecznie ustanowionym ekspertom może łatwo przenieść podobny schemat zaufania na system językowy, którego płynność wypowiedzi przypomina kompetentną komunikację społeczną. Antropomorfizacja nie jest tutaj wyłącznie dziecięcym błędem polegającym na "wierzeniu, że komputer ma uczucia". Jest konsekwencją ludzkiego aparatu społecznego, który odczytuje język, responsywność, pamięć o rozmówcy i dopasowanie emocjonalne jako sygnały obecności kompetentnego partnera. Im sprawniejsza staje się symulacja tych sygnałów, tym większego znaczenia nabiera metapoznawcza zdolność użytkownika do rozdzielania jakości interfejsu komunikacyjnego od wiarygodności treści.

Na tym tle należy przyjrzeć się emocjom i empatii. Empatia ludzka nie jest jednolitą zdolnością; w literaturze rozróżnia się między innymi komponenty afektywne, polegające na współbrzmieniu z cudzym stanem, oraz poznawcze, związane z rozumieniem perspektywy innej osoby. Można trafnie zrekonstruować emocję drugiego człowieka, samemu jej nie doświadczać; można też przeżywać silną reakcję afektywną bez trafnego rozumienia sytuacji drugiej osoby. W związku z tym "syntetyczna empatia" AI wymaga rozłożenia na funkcje. System może identyfikować sygnały emocjonalne, przewidywać reakcje, generować język odbierany jako wspierający i dostosowywać

komunikację bez dowodu na posiadanie własnego stanu afektywnego. Funkcjonalny sukces takiej interakcji jest realny, ale nie rozstrzyga pytania o fenomenologię.

Dryf sprawczości przez outsourcing osądu

Gujral wskazuje na ryzyko dwóch przeciwstawnych redukcji: antropomorfizacji maszyny oraz mechanomorfizacji człowieka. W pierwszym przypadku interpretujemy zachowanie systemu jako dowód na istnienie ludzkopodobnego życia wewnętrznego. W drugim zaś sprowadzamy emocje, relacje i doświadczenie do zbioru mierzalnych sygnałów, jak gdyby pełna predykcja zachowania była tożsama z pełnym wyjaśnieniem podmiotowości. Obie pomyłki wynikają z tego samego błędu metodologicznego: utożsamienia poziomu obserwowalnej funkcji z całym mechanizmem, który ją realizuje.

Ludzka sprawczość nie jest pojedynczą funkcją. Obejmuje ona zdolność formułowania celów, przewidywania konsekwencji, hamowania impulsów, wyboru spośród alternatyw, uczenia się na błędach oraz przypisywania sobie autorstwa decyzji. Nie każda z tych kompetencji musi być rozwinięta w tym samym stopniu. Człowiek może kierować się preferencją, której genezy nie potrafi wyjaśnić; może świadomie przyjąć sugestię z otoczenia lub podjąć decyzję automatycznie, by dopiero później skonstruować dla niej uzasadnienie.

Materiał źródłowy określa jedno z zagrożeń świata algorytmicznego mianem post-rationalization: człowiek może stopniowo stawać się autorem racjonalizacji decyzji zaproponowanych przez system, zamiast pozostać autorem samego procesu wyboru. Teza ta nie zakłada całkowitego zaniku wolnej woli; wystarczy, że część procesów generowania opcji i oceny prawdopodobieństw zostaje trwale przeniesiona poza biologiczny system. Z perspektywy kognitywistycznej można tu wyróżnić trzy poziomy delegacji. Na pierwszym człowiek zleca wykonanie operacji, zachowując cel i kryteria oceny. Na drugim deleguje również generowanie strategii, lecz nadal sam je porównuje. Na trzecim system współtworzy kryteria oceny, selekcjonuje informacje i przedstawia preferowaną opcję w sposób, który wpływa na ludzkie wartościowanie. Im dalej przesuwamy się w tej sekwencji, tym mniej trafna staje się kategoria prostego „narzędzia”.

Nie oznacza to jeszcze powstania autonomicznej AI, lecz wyłonienia się poznawczo współzależnego systemu, w którym ustalenie autorstwa decyzji wymaga analizy całej pętli. To tutaj teoria człowieka jako systemu poznawczego zderza się z problemem kontroli. Formalna kontrola może sprowadzać się do naciśnięcia ostatniego przycisku. Kontrola epistemiczna wymaga jednak rozumienia podstaw rekomendacji, kompetencyjna – zdolności do odrzucenia błędnego wyniku, a temporalna – wystarczającego czasu na refleksję. Kontrola informacyjna zakłada dostęp do alternatywnych

źródeł, natomiast kontrola motywacyjna wymaga, by koszt zakwestionowania systemu nie był tak wysoki, że akceptacja staje się faktycznym ustawieniem domyślnym.

Można zatem pozostawić człowieka „in the loop", a jednocześnie zbudować system, w którym jego obecność jest prawie ceremonialna. Biologiczna kognicja jest szczególnie podatna na taki układ, gdyż naturalnie dąży do oszczędności wysiłku. Jeśli zewnętrzne źródło długotrwale dostarcza trafnych odpowiedzi, doświadczenie zwiększa zaufanie. Jeśli interfejs minimalizuje tarcie, koszt weryfikacji alternatywy rośnie względem kosztu akceptacji. Jeśli system komunikuje się płynnie i personalizuje odpowiedź, uruchamia mechanizmy społeczne. Dodatkowo, pamiętając historię użytkownika, zyskuje dostęp do kontekstu, którego może nie znać pojedynczy ekspert.

Każdy z tych elementów może być indywidualnie korzystny, jednak ich splot tworzy warunki dla tego, co Gujral nazywa dryfem sprawczości: stopniowego outsourcingu osądu powodowanego nie przemocą, lecz wygodą. Ochrona sprawczości nie może zatem sprowadzać się do naiwnego postulatu „myśl samodzielnie". Człowiek zawsze myśli przy pomocy języka, kultury, innych ludzi i narzędzi. Problemem nie jest sama zależność, lecz jej architektura.

Epistemiczne tarcie i granice nadzoru w systemach hybrydowych

Należy zadać pytanie, czy użytkownik widzi źródła informacji, zna poziom niepewności i może łatwo poprosić o alternatywną hipotezę. Czy ma prawo zatrzymać automatyczne działanie, zachowuje dostęp do niezależnej wiedzy i rozumie, które elementy decyzji zostały wygenerowane poza nim? Autonomia w świecie hybrydowym nie wymaga zatem izolacji od AI, lecz dobrze zaprojektowanych mechanizmów epistemicznego tarcia. W tym ujęciu ludzka niedoskonałość staje się parametrem projektowym technologii; system stworzony dla idealnego, racjonalnego operatora byłby bowiem systemem przeznaczonym dla osoby, która nie istnieje.

Human-centered AI powinno uwzględniać ograniczoną uwagę, zmęczenie, efekty ramowania, nadmierne zaufanie do automatyzacji, presję czasu oraz społeczne sygnały wiarygodności. Odpowiedzialność projektanta nie kończy się więc na uzyskaniu wysokiej dokładności modelu. Jeśli system z 95-procentową trafnością komunikuje wyniki w sposób powodujący stuprocentową uległość człowieka, jego ryzyko społeczne nie daje się opisać samą liczbą 95 procent. Potrzebujemy metryk obejmujących jakość kalibracji, możliwość zakwestionowania wyniku, zachowanie kompetencji użytkownika i konsekwencje błędu całego układu.

Z biologicznej natury kognicji wynika dodatkowe ograniczenie: człowiek nie posiada bezpośredniego dostępu do pełnych przyczyn własnych decyzji. Introspekcja jest realną zdolnością, ale nie doskonałym oknem na wszystkie mechanizmy poprzedzające działanie. Procesy percepcyjne, automatyczne skojarzenia, afekt, pamięć i uprzednie nastawienie mogą wpływać na wybór bez pełnej świadomości podmiotu. Późniejsze uzasadnienie często nie jest wierną rekonstrukcją przyczyn, lecz racjonalizacją post-hoc. Z tego powodu pojęcie „human oversight" wymaga ostrożności nawet wtedy, gdy człowiek deklaruje, że świadomie zaakceptował rekomendację. Sam raport użytkownika nie rozstrzyga bowiem, jak duży wpływ na wybór miała konstrukcja interfejsu, kolejność opcji czy wcześniejsza sugestia modelu.

Podobny problem pojawia się po stronie AI, choć z innych przyczyn. Wygenerowane przez model wyjaśnienie własnej odpowiedzi nie musi być transparentnym odczytem rzeczywistych mechanizmów generowania wyniku. W efekcie system hybrydowy może składać się z dwóch podmiotów komunikujących uzasadnienia, które są częściowo rekonstrukcyjne: człowiek może nie mieć pełnego introspekcyjnego dostępu do genezy swojego wyboru, a model może tworzyć przekonujące racjonalizacje bez gwarancji ich wierności wobec wewnętrznej ścieżki przetwarzania. To podwójne ograniczenie wymusza stworzenie zewnętrznej infrastruktury kontroli: danych, eksperymentów, audytu, porównywania hipotez i rygorystycznych procedur.

Dopiero w tym kontekście można prawidłowo zinterpretować rolę intuicji. Nie jest ona nadnaturalnym źródłem wiedzy, lecz w wielu domenach szybkim rozpoznawaniem wzorców, wypracowanym dzięki długiemu doświadczeniu. Ekspert potrafi dostrzec sygnał niewidoczny dla nowicjusza, choć nie zawsze potrafi natychmiast werbalnie wskazać wszystkie przesłanki. Wiarygodność intuicji zależy jednak od struktury środowiska i jakości sprzężenia zwrotnego. W domenach o regularnych zależnościach kompetencje intuicyjne mogą być znakomite; w środowiskach losowych, manipulowanych lub dostarczających niejednoznaczne informacje, poczucie eksperckości bywa mylące. Nie istnieje zatem ogólny „prymat intuicji nad logiką". Istnieją jedynie warunki, w których skumulowane doświadczenie pozwala na szybkie rozpoznanie sytuacji, oraz takie, w których niezbędna jest analityczna kontrola. Korekta ta będzie szczególnie istotna przy omawianiu historii odnalezienia dzieci Huitoto w kolumbijskiej Amazonii.

Notatka interpretuje tę historię jako argument za przewagą doświadczenia i intuicji nad „czystym Systemem 2". W artykule przykład ten może pełnić funkcję antropologiczną i heurystyczną – pokazywać wartość lokalnej wiedzy ekologicznej, praktyk kulturowych i doświadczenia środowiska – ale nie może służyć jako dowód eksperymentalny na istnienie AGE, wyjątkowych kanałów percepcji czy nadrzędność intuicji nad metodami technicznymi. Rozróżnienie między ilustracją a dowodem jest warunkiem zachowania naukowego charakteru całej konstrukcji.

Ludzka kognicja okazuje się więc wielowarstwowym układem adaptacyjnym. Ograniczona racjonalność chroni przed nierealnym wymogiem pełnej optymalizacji; heurystyki umożliwiają szybkie działanie kosztem określonych błędów; uwaga selekcjonuje świat, a pamięć przekształca przeszłość w strukturę przyszłych reakcji. Neuroplastyczność pozwala na uczenie, ale i utrwalanie nawyków; interocepcja i homeostaza zakotwiczają wartościowanie w stanie organizmu; emocje modulują priorytety. Metapoznanie pozwala systemowi oceniać własną pewność, choć samo pozostaje omylne, a poznanie społeczne umożliwia korzystanie z cudzej wiedzy, tworząc jednocześnie podatność na fałszywy autorytet.

Inteligencja jako wynik ucieleśnionej historii

Żaden z tych mechanizmów w izolacji nie definiuje człowieczeństwa. Razem tworzą jednak architekturę, w której inteligencja jest nierozerwalnie związana z ograniczeniem, uczeniem się, cielesnością i historią. Z tej perspektywy kluczowe staje się rozróżnienie trzech pojęć, które w debacie o AI bywają zbyt łatwo utożsamiane: kompetencji, osądu i mądrości. Kompetencja to zdolność rozwiązania określonej klasy problemów. Osąd wymaga już wyboru tego, które problemy należy rozwiązać, jakie informacje uznać za istotne i jak zarządzać ryzykiem w warunkach niepewności. Mądrość zaś – jeśli termin ten ma zachować znaczenie analityczne – odnosi się szerzej do zdolności sytuowania decyzji w długim horyzoncie konsekwencji, wartości, relacji oraz ograniczeń własnej wiedzy. System może posiadać kompetencję bez osądu; człowiek może posiadać oba te elementy, nie dysponując mądrością.

AGI definiowana wyłącznie przez szerokość kompetencji nie rozstrzyga automatycznie kwestii osądu ani mądrości. To tutaj koncepcja AGE znajduje swoje właściwe miejsce. W swojej najciekawszej wersji nie postuluje ona magicznego wyposażenia maszyny w ludzkie emocje, lecz wskazuje, że system zdolny do działania w świecie powinien być oceniany nie tylko przez pryzmat zestawu wykonywanych zadań, ale przede wszystkim przez jego relację z historią własnych działań. Czy pamięta on poprzednie konsekwencje? Czy doświadczenie zmienia jego zachowanie w sposób trwały i generalizowalny? Czy odróżnia zdarzenie statystycznie częste od ważnego dla swoich celów? Czy potrafi rozpoznać własną niepewność lub dysponuje modelem samego siebie, który pozwala przewidywać własne ograniczenia? Czy konsekwencje błędu stają się czymś więcej niż kolejnym rekordem wejściowym?

Są to pytania funkcjonalne, które można badać bez rozstrzygania kwestii świadomości. Jednocześnie analiza biologicznego człowieka pokazuje, dlaczego odpowiedź nie będzie prosta. U nas pamięć, wartościowanie, doświadczenie, ciało i tożsamość są splecione wspólną historią

rozwojową. W systemie technicznym możemy konstruować ich funkcjonalne odpowiedniki oddzielnie: pamięć długoterminową, mechanizm nagrody, sensory, model świata, moduł samooceny czy reguły aktualizacji. Pytanie brzmi, czy głęboka integracja tych elementów może wytworzyć nową jakość poznawczą, czy pozostanie jedynie architekturą symulującą rezultaty procesów biologicznych bez ich doświadczeniowego wymiaru. Odpowiedzi nie należy zakładać z góry ani po stronie biologicznego ekskluzywizmu, ani technologicznego funkcjonalizmu. Najważniejsza lekcja płynąca z biologicznej architektury kognicji jest subtelniejsza niż teza o „wyższości człowieka". Natura nie stworzyła abstrakcyjnego silnika rozumowania, by później dodać do niego emocje, ciało i potrzeby. Inteligencja człowieka wyłoniła się właśnie z systemu, który musiał przetrwać, działać w niepewności, uczyć się przy ograniczonych zasobach, przewidywać innych i zachowywać ciągłość własnej organizacji. To, co z perspektywy klasycznej racjonalności wygląda jak zbiór komplikacji otaczających rozum, z perspektywy biologii jest środowiskiem jego powstania.

Dlatego pytanie nie może brzmieć po prostu: czy AI potrafi robić to samo co człowiek? Musimy zapytać, jaka architektura umożliwia maszynie osiąganie podobnych wyników, które cechy człowieka są w niej zbędne, jakie zostały zastąpione funkcjonalnie, a które okażą się niezbędne dopiero przy przejściu od jednorazowego rozwiązywania problemów do ciągłego działania w realnym świecie. W kolejnej fazie analizy należy zatem opuścić biologiczny system odniesienia i zbadać sztuczną kognicję na jej własnych prawach: nie jako „gorszy mózg", lecz jako historycznie odmienną rodzinę architektur – od uczenia statystycznego i modeli generatywnych, przez multimodalność i modele świata, aż po pamięć, agentowość i mechanizmy samokontroli. Dopiero wtedy okaże się, czy różnica między człowiekiem a AI polega na skali kompetencji, rodzaju architektury, czy właśnie na brakującej osi doświadczenia.

> Być może największy paradoks naszej relacji z technologią polega na tym, że dążymy do stworzenia maszyn idealnie racjonalnych, podczas gdy nasza własna inteligencja zrodziła się właśnie z konieczności radzenia sobie z niedoskonałością. Czy zatem AI, pozbawiona lęku o własne przetrwanie i ciężaru biologicznej historii, będzie w stanie kiedykolwiek naprawdę rozumieć świat, czy jedynie perfekcyjnie go symulować? Ostatecznie to nie skala obliczeń, lecz właśnie nasza kruchość i ucieleśnienie mogą okazać się ostatnim bastionem ludzkiej wyjątkowości.

Inspiracje

[1] "The AI Instinct" Rana Gujral

Rana Gujral to amerykański przedsiębiorca, dyrektor i technolog zajmujący się sztuczną inteligencją, obliczeniami kognitywnymi i przetwarzaniem sygnałów głosowych. Pełnił funkcję dyrektora generalnego Behavioral Signals, firmy produkującej oprogramowanie dla przedsiębiorstw, wykorzystującej uczenie głębokie i przetwarzanie sygnałów mowy do analizy ludzkich emocji, intencji i czynników ryzyka wynikających z interakcji głosowych. Przed pracą w Behavioral Signals, Gujral założył startup TiZE, oferujący oprogramowanie jako usługę w modelu uczenia maszynowego, który został przejęty przez Alchemy, i piastował stanowiska kierownicze w takich firmach jak Logitech i Cricut. Jego praca koncentruje się na komercyjnych i poznawczych obszarach sztucznej inteligencji, interakcji człowiek-maszyna oraz bezpieczeństwa poznawczego. Gujral jest znany z opracowywania koncepcji dotyczących Sztucznego Doświadczenia Ogólnego (AGE) i hybrydowych systemów podejmowania decyzji opartych na sztucznej inteligencji i sztucznej inteligencji, badając, jak inteligencja maszynowa zmienia ludzkie osądy, sprawczość i strategię organizacyjną.

Rana Gujral *is an American entrepreneur, executive, and technologist working in artificial intelligence, cognitive computing, and voice signal processing. He served as the Chief Executive Officer of Behavioral Signals, an enterprise software company utilizing deep learning and speech-signal processing to analyze human emotions, intent, and risk factors from voice interactions. Prior to Behavioral Signals, Gujral founded the machine-learning software-as-a-service startup TiZE, which was acquired by Alchemy, and held executive leadership roles at companies including Logitech and Cricut. His work centers on the commercial and cognitive frontiers of artificial intelligence, human-machine interaction, and cognitive safety. Gujral is known for formulating concepts around Artificial General Experience (AGE) and hybrid human-AI decision-making systems, exploring how machine intelligence reshapes human judgment, agency, and organizational strategy.*

Behavioral Signals

Literatura uzupełniająca

Książki

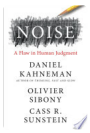
Książki autorów przywoływanych i literatura pokrewna:



[1] "Human Problem Solving"

Herbert Simon

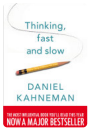
Prentice Hall



[2] "Noise: A Flaw in Human Judgment"

Daniel Kahneman

Hachette UK | ISBN: 9780316451383



[3] "Thinking, Fast and Slow"

Daniel Kahneman

Penguin UK | ISBN: 9781846140556



[4] "Descartes' error : emotion, reason, and the human brain"

Antonio Damasio

Penguin | ISBN: 9781101640135

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Human Problem Solving."

Nick Axten, Allen Newell, Herbert A. Simon

1973 | Contemporary Sociology A Journal of Reviews | DOI: 10.2307/2063712

Źródło

[2] "Verbal reports as data."

K. Anders Ericsson, Herbert A. Simon

1980 | Psychological Review | DOI: 10.1037/0033-295x.87.3.215

Źródło

[3] "Protocol Analysis: Verbal Reports as Data"

M. Venkatesan, K. Anders Ericsson, Herbert A. Simon

1986 | Journal of Marketing Research | DOI: 10.2307/3151491

Źródło

[4] **"Models of Bounded Rationality"**

Herbert A. Simon

1997 | The MIT Press eBooks | DOI: 10.7551/mitpress/4711.001.0001

Źródło

[5] **"Prospect Theory: An Analysis of Decision under Risk"**

Daniel J. Kahneman, Amos Tversky

1979 | Econometrica | DOI: 10.2307/1914185

Źródło

[6] **"Judgment under Uncertainty: Heuristics and Biases"**

Amos Tversky, Daniel J. Kahneman

1974 | Science | DOI: 10.1126/science.185.4157.1124

Źródło

[7] **"Availability: A heuristic for judging frequency and probability"**

Amos Tversky, Daniel J. Kahneman

1973 | Cognitive Psychology | DOI: 10.1016/0010-0285(73)90033-9

Źródło

[8] **"Loss Aversion in Riskless Choice: A Reference-Dependent Model"**

Amos Tversky, Daniel J. Kahneman

1991 | The Quarterly Journal of Economics | DOI: 10.2307/2937956

Źródło

[9] **"Summary of New Challenges to the Rationality Assumption"**

Daniel Kahneman

2013 | Princeton University Press eBooks | DOI: 10.2307/jj.41003669.47

Źródło

[10] **"Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment"**

Amos Tversky, Daniel Kahneman

2008 | Reasoning | DOI: 10.1017/cbo9780511814273.008

Źródło

[11] "Automatic–Heuristic and Executive–Analytic Processing during Reasoning: Chronometric and Dual-Task Considerations"

Wim De Neys

2005 | Quarterly Journal of Experimental Psychology | DOI: 10.1080/02724980543000123

Źródło

[12] "The Smart System 1: evidence for the intuitive nature of correct responding on the bat-and-ball problem"

Bence Bagó, Wim De Neys

2019 | Thinking & Reasoning | DOI: 10.1080/13546783.2018.1507949

Źródło

[13] "Defining System 2 deliberation for dual-process models"

Wim De Neys

2025 | DOI: 10.31234/osf.io/f7436_v1

Źródło

[14] "Humans and LLMs rate deliberation as superior to intuition on complex reasoning tasks"

Wim De Neys, Matthieu Raoelison

2025 | Communications Psychology | DOI: 10.1038/s44271-025-00320-8

Źródło

[15] "How to measure metacognition"

Stephen M. Fleming, Hakwan C. Lau

2014 | Frontiers in Human Neuroscience | DOI: 10.3389/fnhum.2014.00443

Źródło

[16] "Prefrontal Contributions to Metacognition in Perceptual Decision Making"

Stephen M. Fleming, Josefien Huijgen, Raymond J. Dolan

2012 | Journal of Neuroscience | DOI: 10.1523/jneurosci.6489-11.2012

Źródło

[17] "Metacognition: computation, biology and function"

Stephen M. Fleming, Raymond J. Dolan, Chris Frith

2012 | Philosophical Transactions of the Royal Society B Biological Sciences | DOI: 10.1098/rstb.2012.0021

Źródło

[18] **"Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance"**

Marion Rouault, Tricia X. F. Seow, Claire M. Gillan, Stephen M. Fleming

2018 | Biological Psychiatry | DOI: 10.1016/j.biopsych.2017.12.017

Źródło

[19] **"Fear, anger, and risk."**

Jennifer Susan Lerner, Dacher J. Keltner

2001 | Journal of Personality and Social Psychology | DOI: 10.1037/0022-3514.81.1.146

Źródło

[20] **"Beyond valence: Toward a model of emotion-specific influences on judgement and choice"**

Jennifer Susan Lerner, Dacher J. Keltner

2000 | Cognition & Emotion | DOI: 10.1080/026999300402763

Źródło

[21] **"Emotion and Decision Making"**

Jennifer Susan Lerner, Ye Li, Piercarlo Valdesolo, Karim Kassam

2014 | Annual Review of Psychology | DOI: 10.1146/annurev-psych-010213-115043

Źródło

[22] **"Effects of Fear and Anger on Perceived Risks of Terrorism"**

Jennifer Susan Lerner, Roxana M. Gonzalez, Deborah A. Small, Baruch Fischhoff

2003 | Psychological Science | DOI: 10.1111/1467-9280.01433

Źródło

[23] **"Descarte's error : emotion, reason, and the human brain"**

António R. Damásio

1994 | DigitalGeorgetown (Georgetown University Library)

Źródło

[24] **"Insensitivity to future consequences following damage to human prefrontal cortex"**

Antoine Bechara, António R. Damásio, Hanna Damásio, Steven W. Anderson

1994 | Cognition | DOI: 10.1016/0010-0277(94)90018-3

Źródło

[25] "Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala"

Ralph Adolphs, Daniel Tranel, H. Damasio, António R. Damásio

1994 | Nature | DOI: 10.1038/372669a0

Źródło

[26] "Subcortical and cortical brain activity during the feeling of self-generated emotions"

António R. Damásio, Thomas J. Grabowski, Antoine Bechara, Hanna Damasio, Laura L. Boles Ponto

2000 | Nature Neuroscience | DOI: 10.1038/79871

Źródło

Artykuły naukowe autora:

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch02

Źródło

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch03

Źródło

[29] "The AI Instinct"

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.fmatter

Źródło

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch04

Źródło

[31] "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: e2402135-679b-449f-8e8a-22e2860e6f46