Metamind i Supersociety: analiza koncepcji wielopodłożowej kognicji w świetle The AI Instinct Rany Gujrala

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Tekst analizuje koncepcje Metamind oraz Multi-Substrate Person (MSP) w kontekście wizji Rany Gujrala, rozważając możliwość powstania nowej jednostki kognitywnej łączącej biologiczny mózg z AI. Autor przestrzega przed nadmierną ontologizacją tych pojęć, sugerując przejście od spekulacji o „nowym podmiocie" ku ugruntowanej naukowo inteligencji zbiorowej. Przywołując badania nad czynnikiem c Factor oraz analogie do superorganizmów w biologii (np. kolonii mrówek), artykuł dowodzi, że efektywność systemu hybrydowego nie zależy od sumy kompetencji jednostek, lecz od architektury ich relacji i organizacji sprzężeń. Kluczowym wnioskiem jest uznanie Metamindu za system funkcjonalny, który może wykazywać kolektywne zdolności poznawcze bez konieczności posiadania wspólnej świadomości fenomenalnej.

The text analyzes the concepts of Metamind and Multi-Substrate Person (MSP) within the vision of Rana Gujral, considering the possibility of a new cognitive entity combining the biological brain with AI. The author warns against the excessive ontologization of these terms, suggesting a transition from speculations about a "new subject" toward scientifically grounded collective intelligence. Citing research on the c-factor and analogies to superorganisms in biology (e.g., ant colonies), the article demonstrates that the efficiency of a hybrid system does not depend on the sum of individual competencies, but on the architecture of their relations and feedback organization. The key conclusion is the recognition of Metamind as a functional system that can exhibit collective cognitive abilities without the necessity of possessing a shared phenomenal consciousness.

Artykuł stanowi krytyczną analizę koncepcji Metamind i Multi-Substrate Person (MSP), proponując przejście od spekulatywnych wizji fuzji świadomości w stronę rygorystycznej analizy systemowej. Autor stawia tezę, że hybrydowa inteligencja człowiek-AI nie powinna być postrzegana jako nowa, rozproszona „osoba" o wspólnej fenomenologii, lecz jako złożony układ funkcjonalny. Kluczem do efektywności takiego systemu nie jest prosta suma kompetencji jego komponentów, lecz architektura relacji: modularność, mechanizmy orkiestracji (routing kognitywny) oraz zdolność do zarządzania konfliktami między różnymi reprezentacjami świata. Poprzez odróżnienie wielopodłożowości operacyjnej od podmiotowej, tekst redefiniuje Metamind jako dynamiczny system hybrydowy, w którym synergia wynika z precyzyjnego projektu interakcji i interoperacyjności, a nie z magicznej emergencji nowej tożsamości. Tym samym autor sprowadza futurystyczne hipotezy do poziomu mierzalnych parametrów nauki o inteligencji zbiorowej i projektowaniu systemów hybrydowych.

SŁOWA KLUCZOWE

Metamind | Supersociety | wielopodłożowa kognicja | Multi-Substrate Person | Explanation Engine | inteligencja zbiorowa | hybrid intelligence | transactive memory systems | c factor | routing kognitywny | wielopodłożowość operacyjna | synergia człowiek-AI | architektura poznawcza | fenomenalna świadomość

Metamind jako system funkcjonalny a nie nowa osoba

Metamind, wielopodłożowa kognicja i Supersociety: od inteligencji zbiorowej do hipotetycznego podmiotu cywilizacyjnego

Najbardziej spekulacyjny etap rozważań o Hybrydowej Ludzkości zaczyna się w momencie, gdy przestajemy pytać, czy pojedynczy człowiek może korzystać z AI jako rozszerzenia własnej pamięci, percepcji i osądu, a zaczynamy zastanawiać się, czy całe układy ludzi, agentów sztucznych, instytucji, sieci komunikacyjnych i modeli świata mogą stanowić nową jednostkę kognitywną. Gujral opisuje ten kierunek za pomocą pojęć Metamind, Multi-Substrate Person, Supersociety oraz Explanation Engine. Według tej wizji biologiczny człowiek wnosi do hybrydy system wartości, doświadczenie, intuicję i ucieleśniony sens, podczas gdy AI dostarcza skalowalności obliczeniowej i

zdolności analizy ogromnych zbiorów danych. W rezultacie proces poznawczy miałby przestać kończyć się "na czaszce" i przepływać przez kolejne warstwy cyfrowe, tworząc emergentnego agenta funkcjonalnego. Jest to najmocniejsza ontologicznie część materiału i właśnie dlatego wymaga największej dyscypliny naukowej.

Z faktu, że rozproszony system może wykonywać funkcje poznawcze, nie wynika jednak, że stanowi on jedną osobę, posiada wspólną świadomość czy jeden interes moralny. Sam Gujral przedstawia "multi-substrate minds" jako element końcowej wizji The AI Instinct, a w fragmencie książki z sierpnia 2026 roku określa metamind lub Multi-Substrate Person jako hipotetyczne "ja" rozciągnięte pomiędzy biologicznym i obliczeniowym substratem, którego pamięć, uwaga oraz proces myślenia funkcjonowałyby po obu stronach tej granicy. Należy jasno zaznaczyć, że MSP nie jest obecnie ustaloną kategorią nauk kognitywnych ani empirycznie potwierdzonym typem podmiotu.

Jest to propozycja futurologiczno-filozoficzna. Można ją jednak potraktować jako eksperyment pojęciowy, który zmusza do sprecyzowania warunków, jakie musiałby spełniać system rozproszony, abyśmy mogli mówić o czymś więcej niż tylko o efektywnej współpracy wielu elementów. Bezpieczniejszym punktem wyjścia nie jest zatem Multi-Substrate Person, lecz znacznie lepiej ugruntowana inteligencja zbiorowa.

Ludzie od dawna tworzą systemy wykonujące zadania przekraczające możliwości pojedynczego uczestnika. Uniwersytet, rynek, państwo, przedsiębiorstwo czy wspólnota naukowa posiadają własną pamięć, procedury selekcji informacji, specjalizację funkcjonalną i mechanizmy koordynacji. Nie oznacza to, że uniwersytet posiada fenomenalną świadomość, lecz że na poziomie grupowym można identyfikować własności poznawcze, których nie da się zredukować do kompetencji jednej osoby. Klasyczne badania Woolley, Chabris, Pentlanda, Hashmi i Malone'a dostarczyły empirycznego dowodu na to, że wydajność grup w różnych zadaniach można opisać za pomocą statystycznego czynnika zbiorowej inteligencji, określanego jako c factor. W badaniach obejmujących 699 osób czynnik ten był słabo powiązany ze średnią lub maksymalną inteligencją członków; korelował natomiast m.in. ze społeczną wrażliwością uczestników oraz bardziej równomiernym udziałem w rozmowie.

Wynik ten ma zasadnicze znaczenie: inteligencja zespołu nie jest prostą sumą inteligencji jednostek, lecz zależy od architektury relacji pomiędzy nimi. Właśnie dlatego koncepcję Metamind należy analizować przede wszystkim jako problem organizacji sprzężeń, a nie pytanie o liczbę inteligentnych elementów w systemie. Można połączyć tysiąc ekspertów i tysiąc zaawansowanych modeli AI, a mimo to uzyskać układ chaotyczny, podatny na grupowe myślenie, konflikty interesów czy przeciążenie komunikacyjne. Z drugiej strony właściwie zorganizowana grupa o ograniczonej

kompetencji może wytworzyć zadziwiająco skuteczne zachowanie zbiorowe. Emergencja nie polega tu na magicznym pojawieniu się wyższej inteligencji, lecz na tym, że reguły interakcji generują własności makroskopowe, których nie posiada żaden pojedynczy komponent.

Analogiczny model dostarczają badania organizmów społecznych. Kolonie mrówek, pszczół czy termitów są w biologii analizowane jako superorganizmy, ponieważ ich jednostki realizują wyspecjalizowane funkcje, a cały układ przejawia kolektywną regulację, selekcję miejsc i formy zbiorowego podejmowania decyzji. Przegląd Sasakiego i Pratta pokazuje, że kolonie owadów wykazują analogie do indywidualnych systemów poznawczych w kwestiach takich jak kompromis szybkość-dokładność czy racjonalność zbiorowa. Jednocześnie te same mechanizmy pozytywnego sprzężenia, które umożliwiają osiąganie konsensusu, mogą amplifikować błędy i prowadzić do zbiorowej pomyłki.

Analogia superorganizmu jest produktywna pod warunkiem, że nie zostanie zamieniona w antropomorfizację. Kolonia może posiadać kolektywną funkcję poznawczą bez konieczności przyjmowania jednego pola świadomości przez wszystkie owady. Podobnie Supersociety człowiek-AI może wykazywać zdolność wspólnego rozpoznawania problemów i koordynowania działania, bez podstaw do twierdzenia, że społeczeństwo jako całość "czuje" lub doświadcza siebie jako jedna osoba.

Rozróżnienie między wielopodłożowością operacyjną a podmiotową

Rozróżnienie pomiędzy jednością funkcjonalną a fenomenalną musi pozostać nieprzekraczalnym rygorem dalszego wywodu. Gujral idzie znacznie dalej, przedstawiając Multi-Substrate Person jako formę rozproszonej sprawczości i próbując wesprzeć tę intuicję między innymi badaniami Mike'a Levina nad bioelektryczną organizacją tkanek. Właśnie w tym miejscu konieczna jest korekta.

Z badań nad bioelektrycznymi mechanizmami regulacji morfogenezy, zachowaniem komórek czy plastycznością organizmów nie wynika bowiem, że ludzka osoba może zostać rozciągnięta na krzemowy substrat, ani że pamięć osobowa jest substratowo niezależna. Dane te mogą wspierać znacznie ostrożniejszą tezę: organizacja biologiczna wykorzystuje wielopoziomowe mechanizmy informacji i regulacji, których nie należy redukować wyłącznie do aktywności pojedynczych neuronów. Przejście od tego wniosku do koncepcji MSP wymaga szeregu dodatkowych założeń filozoficznych i empirycznych.

Pojęcie „substratu” bywa samo w sobie mylące. W informatyce algorytm można realizować na różnych urządzeniach, o ile zachowane zostaną odpowiednie operacje formalne; nie wiadomo jednak, czy wszystkie własności kognicji – a zwłaszcza świadomość fenomenalna – posiadają analogiczną niezależność. Jak wykazano wcześniej, funkcjonalizm skłania do uznania takiej możliwości, natomiast stanowiska biologicznie zorientowane zwracają uwagę na ucieleśnienie, metabolizm, homeostazę oraz ciągłą fizyczną dynamikę i historię rozwoju.

Multi-substrate cognition w słabym znaczeniu jest więc banalnie możliwa – człowiek myśli z pomocą mózgu, notatek, komputera i AI. Multi-substrate consciousness pozostaje natomiast otwartą hipotezą. Proponuję zatem wyróżnić trzy poziomy wielopodłożowości.

Pierwszy to wielopodłożowość operacyjna: różne substraty uczestniczą w wykonywaniu jednego zadania (np. człowiek, model AI i baza danych wspólnie tworzą raport). Drugi poziom to wielopodłożowość konstytutywna funkcjonalnie: usunięcie jednego elementu zasadniczo zmienia zdolność całego systemu, a komponent zewnętrzny staje się trwałą częścią sposobu działania konkretnej osoby lub instytucji. Trzeci, najsilniejszy poziom oznaczałby wielopodłożowość podmiotową – istnienie jednego fenomenalnego „ja” realizowanego jednocześnie w biologii i infrastrukturze sztucznej.

Pierwszy poziom istnieje. Drugi staje się technicznie i społecznie realny. Trzeci pozostaje spekulacją. Takie stopniowanie pozwala zachować wartość intuicji Gujrala, nie zamieniając jej w twierdzenie wykraczające poza dowody.

Metamind jako system funkcjonalny a nie nowa osoba

Metamind może być rzeczywistą kategorią systemową, zanim stanie się – o ile kiedykolwiek stanie się – kategorią osobową. Zespół człowiek-AI może posiadać pamięć, zdolność uczenia, mechanizmy selekcji uwagi, planowanie, korektę błędów i system norm. To wystarcza, aby analizować go jako złożony układ poznawczy, lecz nie wystarcza, by przyznać mu doświadczenie fenomenalne albo tożsamość osobową. Współczesna teoria Hybrid Intelligence zmierza właśnie w stronę analizy całego zespołu, a nie modelu traktowanego w izolacji. Badanie Dell'Anny i współautorów z 2026 roku proponuje ocenę zespołów hybrydowych przy pomocy takich własności jak boundedness, interdependence, competency, purposefulness, initiative, normativity i effectiveness. Autorzy podkreślają, że Hybrid Intelligence oznacza takie łączenie inteligencji człowieka i maszyny, które zwiększa ludzkie możliwości i pozwala realizować cele nieosiągalne dla każdego komponentu osobno.

Już sama lista tych własności pokazuje, jak można naukowo rozłożyć źródłowy Metamind na elementy podlegające empirycznej ocenie. Pierwszym z nich jest boundedness, czyli zdolność określenia granicy systemu – problem często niedoceniany w popularnych wizjach rozszerzonego umysłu. Jeśli Metamind ma być jednostką analizy, musimy wiedzieć, co do niego należy. Czy prywatny agent użytkownika jest częścią systemu, a wyszukiwarka jedynie narzędziem? Czy baza publiczna dostępna czasowo również stanowi część Metamind? Czy model działający na serwerze producenta pozostaje elementem osoby, skoro producent może go jednostronnie zmienić? W systemach biologicznych granice organizmu są relatywnie czytelne; w cyfrowych mają charakter funkcjonalny, kontraktowy i czasowy. Drugą własnością jest interdependence. Komponent może być pomocny, nie będąc jednocześnie konstytutywnym.

Jeśli odłączenie AI jedynie spowalnia pracę, integracja jest słabsza niż w sytuacji, gdy bez konkretnej pamięci systemowej osoba nie potrafi odtworzyć historii wieloletniego projektu. W miarę wzrostu współzależności rośnie zasadność analizy hybrydy jako jednej jednostki funkcjonalnej, lecz równocześnie zwiększa się znaczenie infrastruktury, prawa do migracji i ciągłości dostępu. Trzecim warunkiem jest wspólny cel. Funkcjonalna jedność zespołu nie powstaje wyłącznie dzięki wymianie informacji; system musi posiadać mechanizm koordynujący działania względem konkretnego kryterium rezultatu. W organizacjach rolę tę pełnią cele, prawo, procedury, budżety i hierarchie. W Metamind problem staje się bardziej złożony, ponieważ różne komponenty mogą mieć odmienne funkcje celu: człowiek może dążyć do prawdy, firma do zysku, agent do wykonania promptu, a instytucja do zgodności regulacyjnej. Wspólna inteligencja nie oznacza automatycznie wspólnego interesu. To jeden z najsłabszych punktów wizji „fuzji” – techniczna integracja nie rozwiązuje konfliktu wartości, a może go wręcz ukrywać. Im bardziej płynnie działa system, tym łatwiej zapomnieć, że jego komponenty zostały stworzone i finansowane przez różne podmioty. Z tego powodu Metamind musi być analizowany nie tylko jako architektura poznawcza, lecz jako konstytucja interesów.

Czwartym warunkiem jest normativity. Zespół hybrydowy funkcjonujący w medycynie, prawie lub administracji nie może optymalizować wyłącznie trafności przewidywania; jego działalność ograniczają prawa, obowiązki, standardy etyczne i procedury. Normy nie są zatem zewnętrznym dodatkiem do inteligencji zbiorowej, lecz częścią samej architektury. Metamind bez warstwy normatywnej byłby co najwyżej systemem optymalizacji, a nie społecznie legitymizowanym podmiotem decyzji. Piąty warunek stanowi zdolność uczenia się na poziomie zbiorowym. Organizacja nie staje się poznawczo dojrzalsza tylko dlatego, że jej pracownicy indywidualnie uczą się na błędach; musi istnieć mechanizm przenoszący wiedzę jednostkową do procedur, pamięci, szkoleń i przyszłych decyzji.

Tutaj powraca funkcjonalne znaczenie AGE. Metamind zaczyna posiadać historię wtedy, gdy doświadczenie jednego elementu może zmienić zachowanie całej sieci także po odejściu tego elementu. W psychologii organizacji odpowiednikiem takiego mechanizmu są między innymi transactive memory systems. Zespół nie musi przechowywać tej samej informacji w każdym umyśle – wystarczy, że jego członkowie wiedzą, „kto wie co", potrafią kierować pytania do właściwych osób i koordynować specjalizację. Współczesne badania podkreślają znaczenie zarówno shared mental models, jak i transactive memory systems jako kluczowych form team cognition. Systematyczny przegląd z 2025 roku wskazuje przy tym na wzajemne sprzężenia pomiędzy poznaniem zespołowym a rozproszonym przywództwem. AI może radykalnie zmienić transaktywną pamięć.

Inteligencja zbiorowa jako architektura routingu i modularności

Tradycyjnie pracownik pamiętał, który kolega specjalizuje się w danej dziedzinie. W systemie hybrydowym agent może automatycznie odnaleźć eksperta, bazę wiedzy, poprzednią decyzję lub adekwatny model. Badania z 2026 roku nad współpracą człowieka z GenAI w społecznościach programistycznych sugerują, że taka synergia zwiększa szerokość i głębokość dostępnej wiedzy zespołów, choć wpływ na koordynację pozostaje niejednoznaczny.

Metamind może zatem rozwijać pamięć nie poprzez centralne „wiedzenie wszystkiego", lecz przez coraz doskonalszą wiedzę o rozmieszczeniu wiedzy. Prowadzi to do rozróżnienia między shared memory a transactive memory. W pierwszym modelu każdy element ma dostęp do tej samej, wspólnej bazy; w drugim specjalizacja pozostaje rozproszona, ale istnieje mechanizm kierowania zapytań do właściwego źródła. Ten drugi model może być bardziej odporny epistemicznie, gdyż nie wymaga homogenizacji całej wiedzy. Supersociety nie musi więc posiadać jednego centralnego modelu świata – może działać jako federacja specjalistycznych reprezentacji połączonych inteligentnym systemem routingu.

Właśnie topologia sieci może okazać się jednym z najważniejszych determinantów przyszłej inteligencji zbiorowej. Najnowsza synteza network science dotycząca hybrydowych grup człowiek-AI podkreśla, że własności kolektywnego poznania zależą od środowiska zadania, struktury połączeń, indywidualnych mechanizmów uwagi i pamięci, systemu bodźców oraz sposobu przepływu informacji. Autorzy zwracają szczególną uwagę na klasyczne napięcia między exploration a exploitation oraz między efektywnością a redundancją, które w zespołach hybrydowych przybierają nowe formy. Exploration-exploitation trade-off jest tutaj kluczowy.

Sieć silnie zintegrowana potrafi błyskawicznie rozpowszechniać dobre rozwiązanie. Jeśli jednak okaże się ono suboptymalne, ta sama szybkość konwergencji może sprawić, że alternatywy zostaną zbyt wcześnie wygaszone. Sieć bardziej modularna pozwala różnym grupom dłużej eksplorować konkurencyjne strategie, choć odbywa się to kosztem wolniejszej koordynacji. Badania nad collective learning pokazują, że wpływ różnorodności kompetencji zależy od złożoności problemu i gęstości sieci. W Supersociety architektura komunikacji jest zatem częścią samej inteligencji.

Pozwala to zrozumieć, dlaczego maksymalna integracja nie jest równoznaczna z maksymalną inteligencją. Popularna wizja superumysłu sugeruje, że im więcej danych i szybsza komunikacja między wszystkimi elementami, tym lepiej. Teoria systemów złożonych podpowiada jednak coś bardziej subtelnego: nadmierna łączność może prowadzić do zbyt szybkiej synchronizacji, korelacji błędów, zaniku lokalnych eksperymentów i przeciążenia wspólną narracją. Efektywny Metamind może wymagać modularności, opóźnień, niezależności i lokalnej autonomii równie mocno, co sprawnej komunikacji.

Biologiczny mózg również nie jest jednorodną pulą, w której każdy neuron komunikuje się ze wszystkimi. Posiada architekturę modularną, hierarchie, specjalizację i mechanizmy selektywnego dostępu. Choć analogia społeczna jest ograniczona, wskazuje ona na ogólną zasadę: inteligencja może wymagać kontrolowanego ograniczenia przepływu informacji, a nie jego absolutnej maksymalizacji. Jest to istotna korekta źródłowej koncepcji Supersociety.

Materiał opisuje ją jako układ, w którym zaufanie opierałoby się nie na interpersonalnych obietnicach, lecz na przejrzystym przepływie informacji w zintegrowanym systemie człowiek-maszyna. Sama przejrzystość nie wystarcza jednak ani do osiągnięcia prawdy, ani zbudowania zaufania – fałszywa informacja może być przesyłana perfekcyjnie transparentnie, a zintegrowany system może błyskawicznie rozprzestrzeniać systematyczny błąd. Supersociety potrzebuje więc nie tylko komunikacji, lecz procedur walidacji, kontroli konfliktów interesów, prawa sprzeciwu i epistemicznej redundancji.

Właśnie dlatego bardziej użyteczne jest porównanie Supersociety z systemem instytucjonalnym niż z pojedynczym mózgiem. Demokracja konstytucyjna nie działa dzięki temu, że wszystkie instytucje posługują się tą samą reprezentacją problemu. Działa raczej dlatego, że sąd, administracja, parlament, media, nauka i społeczeństwo obywatelskie posiadają różne role, kompetencje, normy oraz mechanizmy wzajemnego ograniczania. Z perspektywy kognicji zbiorowej jest to architektura celowo zachowująca konflikt funkcjonalny.

Hybrydowe Supersociety mogłoby więc być tym inteligentniejsze, im lepiej potrafiłoby utrzymywać pluralizm przy jednoczesnej zdolności do działania. Zbyt mała integracja prowadzi do chaosu, zbyt

duża – do monokultury. Najważniejszą własnością systemu staje się dynamiczna równowaga pomiędzy lokalną autonomią a globalną koordynacją. Tę samą zasadę można odnieść do systemów wieloagentowych.

Tysiąc agentów AI nie tworzy automatycznie „superinteligencji". Jeśli wszyscy dysponują podobnym modelem, danymi i funkcją celu, ich liczba zwiększa przepustowość, ale niekoniecznie zdolność poznawczą. Rój staje się interesujący wtedy, gdy występuje specjalizacja, różnorodność, komunikacja, mechanizmy rozstrzygania konfliktów i możliwość adaptacji architektury do zadania. Kolektywna inteligencja może być zatem traktowana jako problem alokacji poznawczej: który element powinien rozwiązać konkretny podproblem? Kiedy należy połączyć wyniki, a kiedy dopuścić konkurencję? Kto powinien otrzymać dostęp do konkretnej pamięci i kiedy centralny agent ma prawo zakończyć spór? W organizacji są to problemy zarządzania; w Metamind stają się problemami architektury poznawczej.

Gujral wyróżnia w tym kontekście role Dyrygenta, Zarządcy, Strażnika i Sztafety. Można je odczytać jako mechanizmy routingu kognitywnego: Dyrygent ustala kolejność i rozdział zadań, Zarządca dynamicznie alokuje zasoby, Strażnik monitoruje anomalie, a Sztafeta określa moment przekazania pracy między różnymi rodzajami agentów.

Władza ukryta w mechanizmach orkiestracji systemu

W Supersociety role te mogłyby zostać częściowo zautomatyzowane, tworząc coś na kształt "systemu operacyjnego" kolektywnego poznania. Pojawia się jednak pytanie: kto steruje tym mechanizmem? Jeśli centralny metapoziom sam jest AI, to rzeczywiste centrum decyzyjne nie spoczywa w agentach wykonujących poszczególne zadania, lecz w samym mechanizmie ustalającym, kto w ogóle otrzymuje problem do rozstrzygnięcia.

Orkiestracja i kompatybilność zamiast jedności poznawczej

Mechanizm ten jest analogiczny do wspomnianej wcześniej epistemologii filtrów. Władza pierwszego rzędu odpowiada na pytanie; władza drugiego rzędu decyduje, kto ma udzielić odpowiedzi. W strukturze organizacji taki meta-agent zyskuje fundamentalne znaczenie. Jeśli to on rozstrzyga, który ekspert zostanie zapytany, które źródło uzyska priorytet, kiedy przypadek wymaga interwencji człowieka, a kiedy może zostać rozstrzygnięty automatycznie, jego funkcja staje się

połączeniem dyspozytora, redaktora i menedżera epistemicznego. To właśnie na tym poziomie mogą utrwalać się systematyczne biasy, nawet jeśli żadna pojedyncza odpowiedź nie jest fałszywa. Metamind wymaga zatem metakognicji systemowej: zdolności do oceny nie tylko samych odpowiedzi, lecz całej architektury rozwiązywania problemów. Czy zadanie skierowano do właściwego agenta? Czy źródła są wystarczająco różnorodne? Czy system konwerguje zbyt szybko? Czy istnieje konflikt interesów lub czy poziom niepewności wymaga eskalacji? Pytania te stanowią odpowiednik ludzkiej metakognicji przeniesionej na poziom organizacji. Może to prowadzić do powstania systemów drugiego rzędu, które nie rozwiązują problemów świata bezpośrednio, lecz obserwują sposób, w jaki robią to inne systemy. Ich rolą byłoby zarządzanie epistemiczną różnorodnością, weryfikacja niezależności źródeł, kontrola uprawnień, detekcja korelacji błędów oraz reorganizacja sieci. Superinteligencja organizacyjna może zatem wymagać nie jednego genialnego modelu, lecz przede wszystkim wysokiej inteligencji orkiestracji. Jest to zbieżne z nowszymi badaniami nad human-AI teams, w których shared situational awareness traktuje się jako dynamiczny proces wzajemnego dostosowywania reprezentacji sytuacji przez ludzi i systemy. Systematyczna synteza opracowana dla środowisk lotniczych opisuje powstawanie wspólnej świadomości sytuacyjnej poprzez integrację informacji, komunikację i wzajemne przewidywanie, ale także przez fazy rozbieżności, negocjacji i naprawy. Dla Metamind jest to kluczowa wskazówka: jedność poznawcza nie polega na stałej identyczności reprezentacji, lecz na zdolności do wykrywania i korygowania ich rozbieżności. Takie ujęcie znacznie lepiej oddaje funkcjonowanie sprawnego zespołu. Chirurg, anestezjolog i pielęgniarka nie muszą posiadać identycznego modelu sytuacji; każdy z nich utrzymuje wiedzę specyficzną dla swojej roli. Wspólna zdolność działania wymaga jednak wystarczającej zgodności w elementach kluczowych oraz możliwości szybkiego ujawnienia konfliktu. Supersociety również nie powinno dążyć do jednego, totalnego obrazu świata, lecz posiadać mechanizmy interoperacyjności pomiędzy obrazami częściowymi. Tutaj ujawnia się istotna różnica między shared mental model a compatible mental models. Pełna wspólność może być ani możliwa, ani pożądana – specjalista potrzebuje reprezentacji, której nie posiada każdy członek zespołu. Kluczowa jest kompatybilność umożliwiająca współpracę. Analogicznie: biologiczny człowiek i AI nie muszą „myśleć tak samo", aby tworzyć skuteczny system. Wartość hybrydy może wynikać właśnie z heterogeniczności ich reprezentacji, co prowadzi do problemu translacji między substratami.

Explanation Engine jako proces krytycznej współpracy

Człowiek operuje językiem, ucieleśnionymi kategoriami, emocjonalnym wartościowaniem i ograniczoną pamięcią roboczą. Model natomiast posługuje się wysokowymiarowymi reprezentacjami, których człowiek nie jest w stanie bezpośrednio odczytać. Aby te systemy mogły

tworzyć jedną funkcjonalną jednostkę, niezbędna jest warstwa translacyjna. Wyjaśnialność staje się zatem czymś więcej niż tylko kwestią audytu – jest interfejsem łączącym różne formy reprezentacji. To właśnie tutaj pierwotny Explanation Engine może zostać przekształcony w rygorystyczną koncepcję systemową. W materiale przedstawiono go jako efekt synergii ludzkiej i sztucznej kognicji, mający zwiększyć moc wyjaśniającą systemu ponad zwykłą zdolność przewidywania.

Nie należy jednak interpretować tego jako automatycznej zdolności hybrydy do poznawania „surowej prawdy natury". Każdy system poznawczy, zarówno biologiczny, jak i sztuczny, pracuje w oparciu o modele i reprezentacje. Explanation Engine powinien być zatem rozumiany jako instytucjonalna architektura produkowania, krytykowania i ulepszania wyjaśnień. Przewidywanie i wyjaśnianie to bowiem odrębne zadania: model może trafnie przewidzieć ruch systemu, nie reprezentując jednocześnie przyczyn w sposób odpowiadający naukowemu wyjaśnieniu. Rozumienie mechanistyczne wymaga identyfikacji struktur, zależności oraz warunków występowania zjawiska. Wyjaśnienie zyskuje szczególną wartość wtedy, gdy pozwala przewidzieć skutki interwencji, rozpoznać przypadek graniczny lub wskazać przyczynę błędu wcześniejszego modelu. W dojrzałym Explanation Engine proces ten obejmowałby kilka etapów: jeden system wykrywa regularność, drugi tworzy hipotezę przyczynową, trzeci wyszukuje kontrprzykłady, człowiek ocenia sens domenowy, kolejny agent projektuje test, a pamięć instytucjonalna konfrontuje rezultat z wcześniejszymi próbami. Nie ma tu miejsca na jeden magiczny moduł „rozumienia" – wyjaśnienie jest produktem cyklu krytycznej współpracy.

AI może w ten sposób radykalnie zwiększyć efektywność nauki. Nie dzieje się to dlatego, że posiada ona dostęp do rzeczywistości bez filtrów, lecz ponieważ potrafi wielokrotnie szybciej eksplorować przestrzeń hipotez, łączyć dyscypliny, analizować dane, proponować eksperymenty i identyfikować anomalie. Człowiek pozostaje w tym układzie niezbędny nie jako romantyczny strażnik intuicji, lecz jako element pluralistycznej architektury obejmującej ucieleśnioną wiedzę, cele normatywne, ocenę znaczenia oraz odpowiedzialność za interwencję. Nie można jednak zakładać, że to człowiek zawsze wnosi najdoskonalszy element wyjaśnienia; w wielu obszarach AI może odkrywać wzorce nieintuicyjne dla ludzkiego umysłu. Prawdziwie hybrydowy Explanation Engine musi zatem dopuszczać dwukierunkową krytykę: człowiek koryguje model, ale model może również ujawnić błąd ludzkiej teorii. Hybrydowość oznacza wzajemną falsyfikowalność, a nie definicyjne uprzywilejowanie jednego z komponentów.

W tym miejscu należy ponownie odrzucić metaforę „eliminowania szumu DMN" czy wizję bezpośredniego dostępu AI do „surowych danych natury". Default Mode Network nie jest jedynie szumem zasłaniającym rzeczywistość, a poznanie całkowicie pozbawione modeli interpretacyjnych nie stanowi realistycznego ideału nauki. Zarówno ludzki mózg, jak i aparatura pomiarowa czy AI

operują poprzez selekcję. Postęp poznawczy polega na tworzeniu lepszych modeli i skuteczniejszych procedur ich krytyki, a nie na usuwaniu wszystkich filtrów. Można wręcz argumentować, że największym wkładem AI do Explanation Engine będzie nie eliminacja ludzkich ograniczeń, lecz zwiększenie liczby kontrolowanych filtrów, przez które można analizować to samo zjawisko. Model statystyczny, symulacja, teoria mechanistyczna, analiza jakościowa i wiedza ekspercka mogą zostać wówczas zestawione jako różne projekcje tego samego problemu.

Słaby Metamind jako dynamiczny system hybrydowy

Ich niezgodność staje się informacją. Wówczas Supersociety nie jest wspólnotą, która „wie wszystko", lecz systemem zdolnym utrzymywać wiele niekompletnych reprezentacji i rozpoznawać momenty, w których zachodzi między nimi konflikt. To epistemicznie dojrzalszy ideał niż jeden centralny supermodel. Metamind należy tu zestawić z pojęciem supermind rozwijanym w MIT Center for Collective Intelligence. Thomas Malone używa tego terminu w odniesieniu do grup ludzi działających razem w sposób inteligentny, co może obejmować również komputery. Misją MIT CCI jest badanie sposobów łączenia ludzi i maszyn tak, aby kolektywnie działać sprawniej niż pojedyncze osoby, grupy czy urządzenia. Choć stanowi to ważny naukowy precedens dla słabszej wersji Metamind, nie należy utożsamiać obu terminów. Supermind Malone'a jest formą zbiorowej inteligencji, a nie twierdzeniem o powstaniu jednej wielosubstratowej osoby.

Można bowiem skonstruować system zdolny lepiej niż jakikolwiek uczestnik przewidywać epidemię, projektować lek czy zarządzać logistyką, a mimo to nie posiadać podstaw do traktowania go jako podmiotu prawnego lub moralnego. Funkcjonalna superinteligencja nie pociąga za sobą osobowości. Co więcej, badania empiryczne nakazują ostrożność wobec znacznie słabszej tezy, jakoby połączenie człowieka i AI automatycznie prowadziło do synergii. Metaanaliza Vaccaro, Almaatouq i Malone'a, obejmująca 106 badań i 370 efektów, wykazała, że zespoły człowiek-AI średnio poprawiały wyniki względem samych ludzi, lecz osiągały rezultaty gorsze niż najlepszy z dwóch komponentów działających osobno. Straty te były szczególnie wyraźne w zadaniach decyzyjnych, podczas gdy większy potencjał synergii obserwowano w obszarach twórczych. To istotny kontrapunkt dla futurystycznej wizji Fused Mind: fuzja sama w sobie nie jest źródłem przewagi. Jeśli współczesne zespoły hybrydowe często nie osiągają synergii, przyszły Metamind nie może być traktowany jako naturalna konsekwencja dodawania kolejnych inteligentnych elementów. Wymagana jest cała nauka hybrid system design. Najnowsza jakościowa metaanaliza identyfikuje 24 powiązane czynniki wpływające na jakość współpracy człowiek-AI, wskazując na nieliniowość zależności między zaufaniem, przejrzystością, kompetencją, charakterem zadania a organizacją interakcji.

Właśnie na tym poziomie znajduje się realistyczny most pomiędzy dzisiejszymi zespołami a futurystycznym Metamind. Warto zatem odwrócić kolejność rozumowania: nie powinniśmy zakładać najpierw powstania Metamindu, by następnie zastanawiać się nad jego własnościami. Należy badać warunki lokalnej synergii i dopiero później pytać, czy ich skalowanie może wytworzyć nowe własności makroskopowe. Pierwszym takim warunkiem jest właściwy podział specjalizacji. Jeśli AI przewyższa człowieka w danej klasie obliczeń, nie ma sensu utrzymywać sztucznej symetrii. Z kolei jeśli człowiek posiada lokalną, kontekstową albo normatywną wiedzę, której model nie potrafi wiarygodnie zrekonstruować, jego rola również nie powinna być dekoracyjna. Inteligencja zbiorowa wymaga świadomości tego, który komponent ma przewagę w konkretnym rodzaju problemu. Drugim warunkiem jest komunikacja niepewności – bez niej elementy sieci nie potrafią prawidłowo ważyć wzajemnych rekomendacji.

Trzecim jest zdolność identyfikacji konfliktu modeli świata. Czwartym – pamięć pozwalająca uczyć się z wcześniejszych błędów systemowych. Piątym – struktura bodźców zapobiegająca ukrywaniu niewygodnej informacji. Szóstym – możliwość reorganizacji sieci w reakcji na nową klasę problemu. Ten ostatni warunek jest szczególnie istotny. Biologiczne systemy i organizacje o wysokiej odporności nie posiadają jednej niezmiennej architektury działania; tworzą nowe połączenia, przekierowują zasoby i reorganizują role. Metamind musiałby być architektonicznie plastyczny. W przeciwnym razie byłby wielkim, ale kruchym systemem.

Można zatem zdefiniować słaby Metamind jako dynamiczny system hybrydowy zdolny do adaptacyjnej alokacji uwagi, pamięci, kompetencji i sprawczości pomiędzy ludźmi a elementami sztucznymi w celu realizacji wspólnych zadań, przy zachowaniu mechanizmów zbiorowego uczenia i korekty błędów. Definicja ta jest mniej spektakularna niż wizja jednej wielosubstratowej osoby, ale łączy się z istniejącymi badaniami nad team cognition, collective intelligence i multi-agent systems. Można również zaproponować wersję średnią, w której system posiada trwałą granicę, własną pamięć historyczną, względnie stabilne cele i zdolność inicjowania działań niezależnie od pojedynczego człowieka. Taki Metamind zaczyna przypominać organizację posiadającą agentową infrastrukturę. Może istnieć przez dziesięciolecia mimo wymiany ludzi i modeli, a jego tożsamość funkcjonalna opiera się na ciągłości instytucjonalnej, nie biologicznej.

> Być może największym wyzwaniem przyszłości nie będzie zatem techniczna fuzja umysłów, lecz sztuka utrzymania niezbędnego dystansu między nimi. Czy w świecie dążącym do totalnej integracji zdołamy zachować modularność i pluralizm, które są jedynymi prawdziwymi gwarantami inteligencji? Prawdziwy superumysł może okazać

> się nie monolitem, lecz harmonijnym chórem głosów, które potrafią ze sobą nie zgadzać się w sposób konstruktywny.

Inspiracje

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

Rana Gujral (ur. 1976) to amerykański przedsiębiorca technologiczny, inwestor i dyrektor zarządzający specjalizujący się w kognitywnej sztucznej inteligencji, inteligencji głosowej i uczeniu maszynowym. Najbardziej znany jest jako były dyrektor generalny Behavioral Signals, firmy zajmującej się sztuczną inteligencją, opracowującej systemy do analizy emocji, intencji i wzorców zachowań w danych dotyczących mowy. Wcześniej Gujral założył startup oprogramowania chmurowego TiZE (przejęty przez Alchemy) i piastował stanowiska kierownicze w działach produktowych w firmach Logitech i Cricut. Jego praca koncentruje się na sztucznej inteligencji opartej na emocjach, przetwarzaniu sygnałów mowy oraz integracji sztucznej inteligencji z procesami podejmowania decyzji przez ludzi. Jako częsty prelegent i felietonista technologiczny w takich publikacjach jak Forbes i TechCrunch, Gujral zgłębia architekturę kognitywnej sztucznej inteligencji, hybrydowe poznanie człowiek-maszyna oraz koncepcję sztucznego doświadczenia ogólnego (AGE) jako kluczowe ścieżki kształtujące przyszłość sztucznej inteligencji.

Rana Gujral *(born 1976) is an American technology entrepreneur, investor, and executive specializing in cognitive artificial intelligence, voice intelligence, and machine learning. He is best known as the former Chief Executive Officer of Behavioral Signals, an enterprise artificial intelligence company developing systems to analyze emotion, intent, and behavioral patterns in speech data. Prior to that, Gujral founded the cloud software startup TiZE (acquired by Alchemy) and held executive product leadership roles at Logitech and Cricut. His work centers on emotion AI, speech signal processing, and the integration of machine intelligence into human decision-making processes. A frequent keynote speaker and technology columnist for publications such as Forbes and TechCrunch, Gujral explores cognitive AI architectures, hybrid human-machine cognition, and the concept of Artificial General Experience (AGE) as critical pathways shaping the future of artificial intelligence.*

Behavioral Signals

Literatura uzupełniająca

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Human Trust in Artificial Intelligence: Review of Empirical Research"

Ella Glikson, Anita Williams Woolley

2020 | Academy of Management Annals | DOI: 10.5465/annals.2018.0057

Źródło

[2] "Evidence for a Collective Intelligence Factor in the Performance of Human Groups"

Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, Thomas W. Malone

2010 | Science | DOI: 10.1126/science.1193147

Źródło

[3] "The role of gender in team collaboration and performance"

Julia B. Bear, Anita Williams Woolley

2011 | Interdisciplinary Science Reviews | DOI: 10.1179/030801811x13013181961473

Źródło

[4] "Collective Intelligence and Group Performance"

Anita Williams Woolley, Ishani Aggarwal, Thomas W. Malone

2015 | Current Directions in Psychological Science | DOI: 10.1177/0963721415599543

Źródło

[5] "Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events"

Daniel J. Simons, Christopher F. Chabris

1999 | Perception | DOI: 10.1068/p281059

Źródło

[6] "Neural mechanisms of general fluid intelligence"

Jeremy R. Gray, Christopher F. Chabris, Todd Samuel Braver

2003 | Nature Neuroscience | DOI: 10.1038/nn1014

Źródło

[7] "Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments"

Laura Thi Germine, Ken Nakayama, Bradley C. Duchaine, Christopher F. Chabris, Garga Chatterjee

2012 | Psychonomic Bulletin & Review | DOI: 10.3758/s13423-012-0296-9

Źródło

[8] "Wearable feedback systems for rehabilitation"

Carl D. Marci, Sung Michael, Alex ‘Sandy’ Pentland

2005 | DOAJ (DOAJ: Directory of Open Access Journals) | DOI: 10.1186/1743-0003-2-17

Źródło

[9] "MIThril 2003: applications and architecture"

Richard W. DeVaul, M. Sung, Jonathan Gips, Alex ‘Sandy’ Pentland

2004 | Proceedings | DOI: 10.1109/iswc.2003.1241386

Źródło

[10] "The ALIVE system: full-body interaction with autonomous agents"

Pattie Maes, Trevor J. Darrell, Bruce M. Blumberg, Alex ‘Sandy’ Pentland

2002 | Computer Animation (CA) | DOI: 10.1109/ca.1995.393553

Źródło

[11] "Electronic Memory Aids for Community-Dwelling Elderly Persons: Attitudes, Preferences, and Potential Utilization"

Jiska Cohen-Mansfield, Michael A. Creedon, Thomas B. Malone, Mark Kirkpatrick, Lisa A. Dutra

2005 | Journal of Applied Gerontology | DOI: 10.1177/0733464804271277

Źródło

[12] "FIELD TEST EVALUATION OF REAR LIGHTING SYSTEMS"

Thomas B. Malone, Mark A. Kirkpatrick, Johannes Kohl, C F Baker

1978

Źródło

[13] "The Art of Successfully Applying Human Systems Integration"

Alexander C. Landsburg, Larry W. Avery, ROBERT BEATON, J. Robert Bost, Carlos A. Comperatore, Thomas B. Malone

2008 | Naval Engineers Journal | DOI: 10.1111/j.1559-3584.2008.00113.x

Źródło

[14] "Development of an Interactive Microprocessor Based Workload Evaluation Model (SIMWAM)"

Mark A. Kirkpatrick, Thomas B. Malone, Phillip J. Andrews

1984 | Proceedings of the Human Factors and Ergonomics Society Annual Meeting | DOI: 10.1177/154193128402800123

Źródło

[15] "Chess as a Model of Collective Intelligence: analyzing a distributed form of chess with piece-wise agency"

Guillermo CAMPITELLI, David Kofman, Michael Levin

2025 | DOI: 10.31219/osf.io/n62jx_v1

Źródło

[16] "Combining energy dynamic simulation and multi-criteria analysis for supporting investment decisions on smart shading devices in office buildings"

Maria Cristina Pinto, Giulia Crespi, Federico Dell'Anna, Cristina Becchio

2022 | Applied Energy | DOI: 10.1016/j.apenergy.2022.120470

Źródło

[17] "EPC Green Premium in Two Different European Climate Zones: A Comparative Study between Barcelona and Turin"

Federico Dell'Anna, Marina Bravi, Carlos Ramiro Marmolejo Duarte, Marta Bottero, Chen Ai

2019 | Sustainability | DOI: 10.3390/su11205605

Źródło

[18] "Supporting public decision process in buildings energy retrofitting operations: The application of a Multiple Criteria Decision Aiding model to a case study in Southern Italy"

Grazia Napoli, Marta Bottero, Giuseppina Ciulla, Federico Dell'Anna, José Rui Figueira

2020 | Sustainable Cities and Society | DOI: 10.1016/j.scs.2020.102214

Źródło

[19] "What Advantages Do Adaptive Industrial Heritage Reuse Processes Provide? An Econometric Model for Estimating the Impact on the Surrounding Residential Housing Market"

Federico Dell'Anna

2022 | Heritage | DOI: 10.3390/heritage5030082

Źródło

[20] "A Test for Evaluating Performance in Human-AI Systems"

Thomas Malone, Michelle Vaccaro, Andres Campero, Jaeyoon Song, Haoran Wen

2023 | DOI: 10.21203/rs.3.rs-2787476/v1

Źródło

[21] "Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements"

Anthony G. Vaccaro, Stephen M. Fleming

2018 | Brain and Neuroscience Advances | DOI: 10.1177/2398212818810591

Źródło

[22] "How large language models can reshape collective intelligence"

Jason W. Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Abdullah Almaatouq

2024 | Nature Human Behaviour | DOI: 10.1038/s41562-024-01959-9

Źródło

[23] "A Test for Evaluating Performance in Human-Computer Systems"

Andres Campero, Michelle Vaccaro, Jaeyoon Song, Hao-Ran Wen, Abdullah Almaatouq

2022 | arXiv.org | DOI: 10.48550/arxiv.2206.12390

Źródło

[24] "The Effects of Group Composition and Dynamics on Collective Performance"

Abdullah Almaatouq, Mohammed Alsobay, Ming Yin, Duncan J. Watts

2023 | Topics in Cognitive Science | DOI: 10.1111/tops.12706

Źródło

Artykuły naukowe autora:

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch02

Źródło

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch03

Źródło

[27] "The AI Instinct"

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.fmatter

Źródło

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch04

Źródło

[29] "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: a389982d-cb81-4b6b-b2d0-42a9724a5ba3