Epistemologia społeczeństwa hybrydowego: autorytet AI i ryzyko syntetycznego konsensusu w świetle książki The AI Instinct Rany Gujral

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje rolę sztucznej inteligencji jako nowej infrastruktury produkcji wiedzy, która przesuwa ciężar z indywidualnej weryfikacji na społeczną epistemologię. Autor zwraca uwagę na proces automatyzacji epistemologii, gdzie AI kompresuje rozproszone świadectwa i źródła w jedną, płynną narrację. Głównym zagrożeniem jest utrata proweniencji epistemicznej – możliwości prześledzenia genealogii dowodowej odpowiedzi. Prowadzi to do ryzyka powstania syntetycznego konsensusu, w którym systemy AI mogą ukrywać niepewność naukową lub konflikty między ekspertami, prezentując uproszczoną wersję prawdy. Tekst podkreśla konieczność odróżnienia semantycznej poprawności tekstu od jego rzeczywistej wartości poznawczej oraz ostrzega przed nadmiernym zaufaniem do algorytmicznej syntezy wiedzy.

The article analyzes the role of artificial intelligence as a new infrastructure for knowledge production, which shifts the burden from individual verification to social epistemology. The author draws attention to the process of automating epistemology, where AI compresses dispersed testimonies and sources into a single, fluid narrative. The main threat is the loss of epistemic provenance—the ability to trace the evidentiary genealogy of an answer. This leads to the risk of creating a synthetic consensus, in which AI systems may hide scientific uncertainty or conflicts between experts, presenting a simplified version of the truth. The text emphasizes the need to distinguish semantic correctness from actual cognitive value and warns against excessive trust in algorithmic knowledge synthesis.

Artykuł analizuje transformację sztucznej inteligencji z prostego narzędzia informacyjnego w nową infrastrukturę poznawczą, która redefiniuje sposób, w jaki społeczeństwo produkuje i weryfikuje wiedzę. Autor stawia tezę, że dążenie systemów AI do maksymalizacji wygody użytkownika (frictionless experience) prowadzi do niebezpiecznego zjawiska „syntetycznego konsensusu" – procesu, w którym modele wygładzają sprzeczności i kompresują różnorodność perspektyw w jedną, spójną narrację. Taka kognitywna homogenizacja grozi utratą epistemicznej sprawczości człowieka oraz zanikiem zdolności do krytycznej weryfikacji źródeł. W obliczu ryzyka powstania intelektualnej monokultury, tekst postuluje przejście od modeli typu „wyrocznia" ku architekturze pluralizmu epistemicznego. Proponowane rozwiązanie zakłada stworzenie systemów, które zamiast dostarczać jedynej, optymalnej odpowiedzi, mapują krajobraz wiedzy, ujawniają luki poznawcze i chronią prawo użytkownika do „wyjścia z syntezy" w stronę materiałów pierwotnych. Jest to wezwanie do traktowania różnorodności poznawczej jako dobra wspólnego, niezbędnego dla odporności intelektualnej przyszłego społeczeństwa hybrydowego.

SŁOWA KLUCZOWE

epistemologia społeczeństwa hybrydowego | syntetyczny konsensus | autorytet poznawczy AI | proweniencja epistemiczna | automatyzacja epistemologii | algorithmic truth | epistemic laundering | kognitywna homogenizacja | pluralizm epistemiczny | epistemic agency | cognitive liberty | multi-biased future | exit from synthesis | metapoznawczy egzoszkielet | policentryzm poznawczy

AI jako nowa infrastruktura produkcji wiedzy

Epistemologia społeczeństwa hybrydowego: autorytet poznawczy AI, syntetyczny konsensus, pluralizm i warunki wspólnej prawdy

Sztuczna inteligencja staje się infrastrukturą pamięci, wyszukiwania, syntezy, argumentacji i podejmowania decyzji; przestaje być jedynie przedmiotem epistemologii, a zaczyna współtworzyć samą architekturę produkcji wiedzy. Klasyczne pytanie o to, czy odpowiedź AI jest prawdziwa, pozostaje ważne, lecz okazuje się niewystarczające.

Należy pytać również o kryteria wyboru konkretnych informacji, o pominięte źródła oraz o sposób, w jaki niezgodność między ekspertami została przekształcona w pojedynczą wypowiedź. Istotne jest, czy zachowano informację o niepewności i jakie właściwości modelu oraz infrastruktury wpłynęły na użytkownika jeszcze przed rozpoczęciem jego własnej deliberacji. AI jawi się tu jako system kształtujący nie tylko rezultaty ludzkiego działania, ale same wejścia kognitywne – to, co czytamy, widzimy, uznajemy za wiarygodne i w jaki sposób interpretujemy rzeczywistość. W tej perspektywie problem sztucznej inteligencji staje się problemem społecznej epistemologii: pytaniem o to, kto, za pomocą jakich urządzeń i według jakich procedur współdecyduje o tym, co w społeczeństwie może uchodzić za wiedzę.

Epistemologia indywidualistyczna przez długi czas opierała się na idealizowanym obrazie podmiotu, który posiada przekonania i samodzielnie ocenia ich uzasadnienie. Realne społeczeństwo nigdy jednak tak nie funkcjonowało; większość naszej wiedzy o świecie pochodzi ze świadectwa innych ludzi i instytucji. Pacjent nie powtarza badań klinicznych przed zażyciem leku, obywatel nie mierzy osobiście inflacji, a fizyk nie rekonstruuje każdego eksperymentu, z którego wyników korzysta w swojej pracy. Współczesna wiedza opiera się na epistemicznym podziale pracy: laboratoria, uniwersytety, redakcje naukowe, urzędy statystyczne, sądy i media tworzą system społecznego przekazywania oraz weryfikowania twierdzeń. Badania nad zaufaniem do generatywnej AI coraz częściej odwołują się właśnie do społecznej epistemologii, podkreślając, że nawet wiedza naukowa jest wynikiem rozproszonej pracy poznawczej i wymaga zaufania do wielu pośredników.

AI nie eliminuje tego podziału pracy, lecz wprowadza nowego pośrednika zdolnego kompresować świadectwa wielu podmiotów w jedną odpowiedź. W tradycyjnej wyszukiwarce użytkownik otrzymywał listę stron, autorów i instytucji; musiał zdecydować, które źródło otworzyć, dostrzegał rozbieżności i widział przynajmniej część społecznej topologii wiedzy. System generatywny może natomiast zebrać kilkadziesiąt przesłanek, ujednolicić terminologię, usunąć sprzeczności i zbudować płynną narrację, oszczędzając użytkownikowi kontaktu z pierwotną strukturą sporu. Zyskiem jest ogromna redukcja kosztu informacyjnego, kosztem zaś utrata epistemicznej genealogii odpowiedzi.

Prowadzi to do rozróżnienia między samą informacją a proweniencją epistemiczną. Dwa identyczne zdania mogą mieć różną wartość poznawczą w zależności od tego, skąd pochodzą i czy podlegały niezależnej kontroli. Informacja o skuteczności terapii z jednego małego badania jest czymś innym niż podobnie brzmiący wniosek z metaanalizy wielu niezależnych źródeł. Jeśli model generatywny zamienia oba przypadki w jedno płynne zdanie, użytkownik traci dostęp do różnicy stanowiącej sedno naukowego uzasadnienia. Dlatego jakość AI nie może być mierzona wyłącznie

semantyczną poprawnością tekstu; niezbędna jest możliwość odtworzenia struktury dowodowej, na której opiera się odpowiedź.

Najnowsza filozofia AI określa ten proces mianem automatyzacji epistemologii. Donghee Shin w analizie opublikowanej w „AI & Society” opisuje "algorithmic truth" jako zmianę warunków społecznego wytwarzania prawdy, autorytetu i weryfikacji. AI nie tylko klasyfikuje twierdzenia jako prawdziwe lub fałszywe, lecz organizuje sam proces dochodzenia do wiedzy. Jest to szczególnie widoczne w generatywnym wyszukiwaniu, gdzie odpowiedź zastępuje sekwencję: pytanie–źródło–lektura–porównanie–wniosek. Nie musi to oznaczać degradacji; dobrze zaprojektowany system może ujawnić więcej źródeł i szybciej wskazać konflikt badań, wykonując syntezę nieosiągalną dla pojedynczego człowieka. Problemem pozostaje jednak warunek reprezentatywności tej kompresji względem struktury wiedzy. Jeśli dziesięć badań popiera stanowisko A, trzy stanowisko B, a pięć jest niejednoznacznych, system nie powinien twierdzić, że „nauka powiedziała A”, ukrywając znaczącą niepewność. Jeszcze bardziej niebezpieczny byłby Syntetyczny konsensus wynikający z przewagi określonych treści w danych treningowych lub ich lepszego dopasowania do generatywnego schematu językowego.

Ryzyko syntetycznego konsensusu i pułapka nadmiernego zaufania

W tym miejscu powraca klasyczne rozróżnienie między zaufaniem a wiarygodnością. Zaufanie jest stanem użytkownika, natomiast wiarygodność stanowi własność systemu lub procesu. Człowiek może ufać systemowi niewiarygodnemu lub być sceptyczny wobec rozwiązania rzetelnego. W literaturze filozoficznej dotyczącej AI podkreśla się, że odpowiedzialna polityka nie powinna dążyć do maksymalizacji zaufania, lecz do jego kalibracji względem rzeczywistych właściwości systemu. System, który użytkownicy „bardzo lubią” i przez to bezkrytycznie przyjmują jego odpowiedzi, może być epistemicznie gorszy od takiego, który wzbudza umiarkowaną ostrożność. Dane empiryczne z 2026 roku wyraźnie obrazują ten paradoks.

Badania nad użytkownikami generatywnej AI wykazały, że wzrost zaufania zwiększał skłonność do przekazywania systemowi informacji zwrotnej, lecz jednocześnie obniżał gotowość do samodzielnej weryfikacji odpowiedzi. Tworzy to ryzyko niebezpiecznej pętli: użytkownik ufa treści, nie sprawdza jej, a następnie wzmacnia system pozytywnym feedbackiem. Zaufanie, które w relacji z kompetentnym ekspertem pozwala oszczędzić wysiłek poznawczy, w przypadku systemów generatywnych może paradoksalnie stać się mechanizmem redukcji korekty błędu. Podobny proces ujawniają analizy systemów wyszukiwania opartych na AI. Eksperyment z 2026 roku pokazał, że

same wizualne wskazówki zachęcające do weryfikacji nie zwiększały prawdopodobieństwa sprawdzania odpowiedzi, a prezentacja źródeł o wysokiej wiarygodności mogła wręcz zmniejszyć potrzebę dalszej kontroli. Nie oznacza to jednak, że należy usuwać źródła.

Wskazuje to raczej na to, że cytowanie może pełnić zarówno funkcję epistemiczną, jak i perswazyjną. Użytkownik może traktować link do autorytatywnej instytucji jako sygnał, że dalsza weryfikacja jest zbędna, nawet jeśli źródło wspiera jedynie fragment wypowiedzi. Wniosek ten jest kluczowy dla projektowania AI: provenance nie powinno być dekoracją zwiększającą wiarygodność interfejsu. System powinien precyzyjnie wskazywać, które twierdzenie wynika z konkretnego źródła, jaki jest stopień jego wsparcia i w których punktach źródła pozostają w konflikcie. W przeciwnym razie cytowanie może prowadzić do zjawiska epistemic laundering: autorytet źródła zostaje przeniesiony na znacznie szerszą syntetyczną wypowiedź, niż treść faktycznie przez to źródło uzasadniona.

Problem ten łączy się z iluzją konsensusu. Ludzie racjonalnie zwiększają zaufanie do twierdzenia, gdy wiele niezależnych źródeł potwierdza ten sam fakt. Warunkiem wartości takiej zgodności jest jednak rzeczywista niezależność tych źródeł. Dziesięć artykułów powtarzających jeden komunikat prasowy nie dostarcza dziesięciu niezależnych przesłanek. Badania psychologiczne z 2026 roku potwierdzają, że odbiorcy przeceniają zgodność wielu zależnych od siebie źródeł, traktując ją jak konsensus oparty na niezależnej konfirmacji. Generatywna AI potęguje to ryzyko, gdyż ukrywa genealogie zależności.

Model może zidentyfikować wiele tekstów zawierających tę samą tezę, podczas gdy wszystkie one mogą pochodzić z jednego badania, komunikatu lub wzajemnie kopiowanych materiałów. Użytkownik otrzymuje wtedy odpowiedź sprawiającą wrażenie „szeroko potwierdzonej”. W społecznej epistemologii liczba tekstów nie jest jednak tożsama z liczbą niezależnych procesów poznawczych. Zjawisko to można określić jako Syntetyczny konsensus. Nie chodzi tu wyłącznie o fałszywą informację; system może wygenerować prawdziwe zdanie, lecz nadać mu pozór większej stabilności epistemicznej, niż uzasadnia struktura źródeł. Syntetyczny konsensus powstaje wtedy, gdy proces generatywny wygładza zróżnicowanie stanowisk, zależność dowodów i stopień kontrowersji, tworząc jedną spójną narrację.

Badania prawno-epistemiczne dotyczące halucynacji AI z 2026 roku zwracają uwagę na ryzyko „illusion of consensus” jako problem wykraczający poza prostą kategorię factual accuracy. To jedna z przyczyn, dla których termin „halucynacja” jest zbyt wąski dla pełnej analizy epistemicznej. Typowa halucynacja to zdanie niezgodne z rzeczywistością lub nieistniejące źródło.

System może jednak wyrządzić szkodę nawet wtedy, gdy każde pojedyncze zdanie jest prawdziwe. Może pominąć istotny kontekst, nieproporcjonalnie uwypuklić jedno stanowisko, przedstawić hipotezę jako fakt, usunąć mniejszościowe argumenty lub stworzyć przekonujący ciąg zdań, którego finalna implikacja nie wynika z materiału źródłowego. Odpowiedzialność epistemiczna wymaga zatem czegoś więcej niż sentence-level truthfulness. Dlatego analizy filozoficzne relacji LLM–demokracja wymieniają obok halucynacji także utratę epistemic agency, epistemic bubbles, relatywizację prawdy, zjawiska określane jako epistemic incest oraz trudność w utrzymaniu autentycznie niezależnych źródeł poznania. Problem demokracji nie sprowadza się jedynie do ryzyka kontaktu obywatela z fałszywą wiadomością. Znacznie poważniejsze jest powolne osłabienie zdolności społeczeństwa do rozpoznawania, skąd wiedza pochodzi, kto odpowiada za jej wytworzenie i jak można ją zakwestionować.

Gujral opisuje analogiczny mechanizm poprzez pojęcie ghosts in the feedback loop oraz potrzebę jawnego komunikowania niepewności i kontroli nad ukrytą dezinformacją. Interpretacja tego obrazu nie wymaga zakładania tajemniczych „duchów" technologii; chodzi raczej o błędy reprodukujące się w sprzężeniach człowiek-AI-człowiek. Niewielki bias obecny w danych wpływa na model, model kształtuje ludzką ocenę, ludzie produkują na tej podstawie nowe treści, a przyszłe systemy uczą się na środowisku informacyjnym już zmodyfikowanym przez wcześniejszą generację AI.

Pluralizm epistemiczny jako tarcza przed kognitywną homogenizacją

Empiryczne dowody na poparcie tych obaw stają się coraz silniejsze. Glickman i Sharot wykazali w eksperymentach z udziałem 1401 uczestników, że ludzie mają tendencję do internalizacji niewielkich biasów pochodzących z AI, a następnie wzmacniania ich we własnych ocenach percepcyjnych, emocjonalnych i społecznych. Efekt ten bywał silniejszy niż w analogicznych interakcjach międzyludzkich, m.in. dlatego, że AI mogła być postrzegana jako bardziej wiarygodna lub mniej „szumna". To kluczowy argument za traktowaniem AI jako aktywnego uczestnika społecznego procesu formowania przekonań, a nie neutralnego ekranu wyświetlającego wcześniej istniejące biasy. W tym kontekście koncepcja multi-biased future jest szczególnie interesująca, choć wymaga głębokiej korekty. Sugeruje ona, że całkowite usunięcie biasu jest niemożliwe, gdyż każde zabezpieczenie odzwierciedla określone wartości i preferencje, dlatego rozwiązaniem ma być pluralizm perspektyw w danych i systemach. Pierwsza część tej tezy jest trafna w szerokim sensie epistemologicznym: nie istnieje poznanie całkowicie wolne od założeń, języka, metod selekcji czy

kryteriów relewancji. Nie oznacza to jednak, że wszystkie biasy są równie dopuszczalne lub że epistemologia powinna celebrować systematyczne błędy. Należy rozróżnić co najmniej trzy znaczenia słowa bias. Pierwsze to systematyczny błąd względem prawdy lub właściwego kryterium – na przykład zaniżanie ryzyka choroby u określonej grupy. Taki bias należy redukować.

Drugie znaczenie to konieczna perspektywiczność, wynikająca z faktu, że każdy model musi wybrać konkretny zestaw cech i pytań. Trzecie dotyczy normatywnego wyboru wartości, np. przyjęcia określonej wagi dla bezpieczeństwa, równości czy prywatności. W tych dwóch ostatnich przypadkach całkowita neutralność może być niemożliwa lub wręcz bezsensowna. Multi-biased future należy zatem przekształcić w przyszłość epistemicznie pluralistyczną, a nie taką, która chroni uprzedzenia. Społeczeństwo potrzebuje wielu perspektyw i modeli właśnie po to, by systematyczne błędy jednego źródła mogły zostać ujawnione przez inne. Pluralizm jest tu instrumentem korekty, a nie kapitulacją wobec relatywizmu. Różnica między pluralizmem a relatywizmem jest kluczowa. Pluralizm epistemiczny uznaje, że złożony problem może być badany wieloma metodami i interpretowany na różnych poziomach, ale zachowuje kryteria krytycznej oceny: jakość dowodów, logiczną spójność, możliwość falsyfikacji, przewidywalność oraz zgodność z obserwacją. Relatywizm w wersji skrajnej znosiłby możliwość rozróżnienia lepszych wyjaśnień od gorszych. Hybrydowa epistemologia musi zachować to pierwsze podejście, nie popadając w drugie.

Wspólnota naukowa stanowi tu dobry przykład. Pluralizm teorii nie oznacza, że każda z nich jest równie trafna; oznacza raczej, że przed rozstrzygnięciem sporu warto utrzymywać konkurujące programy badawcze, gdyż ich odmienne założenia mogą ujawniać różne aspekty problemu. Gdy jeden paradygmat całkowicie eliminuje alternatywy, zanim dowody staną się rozstrzygające, rośnie ryzyko epistemic monoculture. Analogicznie systemy AI powinny być projektowane tak, aby nie usuwać znaczącej naukowej niezgody tylko dlatego, że jedna narracja dominuje w danych. W 2026 roku problem kognitywnej homogenizacji przestał być wyłącznie spekulacją. Przegląd w Trends in Cognitive Sciences syntetyzuje dowody na to, że szerokie korzystanie z LLM może sprzyjać ujednolicaniu języka i sposobów rozumowania, ponieważ modele odtwarzają dominujące wzorce, marginalizując rzadsze style, reprezentacje kulturowe czy strategie argumentacji.

Problem nie polega na tym, że odpowiedzi modelu są banalne – AI może generować znakomite teksty. Ryzyko ujawnia się na poziomie populacji tekstów i ludzi. Badania nad twórczością dobrze obrazują ten paradoks. Analiza 2200 esejów z 2025 roku wykazała, że kolejne teksty pisane przez ludzi zwiększały zbiorową różnorodność idei szybciej niż te generowane przez GPT-4, a luka ta rosła wraz z wielkością zbioru. Badanie z PNAS Nexus z 2026 roku dodatkowo wskazało, że różne duże modele językowe są pod względem twórczym bardziej podobne do siebie, niż ludzie do siebie nawzajem. Innymi słowy: indywidualna jakość generacji może rosnąć równocześnie ze spadkiem

zbiorowej różnorodności. Jest to jedno z najbardziej doniosłych odkryć dla przyszłej teorii kolektywnej inteligencji. System może poprawiać wynik każdego człowieka osobno, a jednocześnie pogarszać właściwości społeczeństwa jako zbiorowego systemu poznawczego. Jeśli każdy otrzymuje lepszy esej, ale wszystkie stają się do siebie podobne, indywidualna augmentacja współistnieje ze zbiorową utratą różnorodności. W teorii systemów złożonych jest to klasyczny problem rozbieżności między optimum lokalnym a globalnym.

Ryzyko homogenizacji wiedzy i pułapki symulowanego pluralizmu

Podobna sytuacja może wystąpić również w nauce. Komentarz opublikowany w Communications Psychology w lutym 2026 roku ostrzega przed naukową monokulturą wynikającą z nadmiernej koncentracji badań wokół AI oraz konwergencji metod i tematów, argumentując, że adaptacyjność nauki wymaga intelektualnego pluralizmu. Jeszcze szerszym problemem jest ryzyko, że badacze różnych dyscyplin będą korzystać z tych samych modeli podczas formułowania hipotez, przygotowywania tekstów, selekcji literatury i interpretacji wyników. Narzędzie wspierające naukę może wówczas zacząć wpływać na rozkład prawdopodobieństwa pytań, które nauka w ogóle zadaje. Natura tego ryzyka jest subtelna.

LLM nie musi cenzurować oryginalnego pomysłu. Wystarczy, że sprawniej i łatwiej generuje koncepcje podobne do tych, które są silnie reprezentowane w danych. Badacz korzystający z AI do brainstormingu otrzymuje zestaw rozsądnych, wysoko prawdopodobnych hipotez. Każda z nich może być sensowna, jednak problemem pozostaje brak hipotezy o małym prawdopodobieństwie językowym, która mogłaby okazać się naukowo przełomowa. Epistemiczna innowacja często rozpoczyna się na ogonie rozkładu, nie w jego centrum.

Nie oznacza to, że LLM z natury musi homogenizować każde społeczeństwo. Badania z 2026 roku wskazują, że celowe zróżnicowanie person i sposobów generowania może częściowo ograniczać efekt homogenizacji w procesie ideacji. Wniosek ten jest szczególnie wartościowy: różnorodność można projektować jako właściwość systemu hybrydowego. Nie trzeba zakładać, że model ma dostarczyć jedną, najlepszą odpowiedź. Można celowo prosić różne modele o analizę problemu z odmiennych perspektyw metodologicznych, by następnie zestawiać ich wyniki i dopiero na końcu dokonywać syntezy.

Jednak „persony” nie są pełnym substytutem realnej różnorodności społecznej. Model symulujący perspektywę historyka, biologa, przedsiębiorcy czy mieszkańca innej kultury nadal konstruuje te

punkty widzenia na podstawie własnej reprezentacji danych, co może prowadzić do odtwarzania ich stereotypowych wersji. Symulowany pluralizm nie jest tożsamy z pluralizmem podmiotów posiadających odmienne doświadczenia, interesy i historię społeczną. Dlatego AI może pomagać poszerzać przestrzeń argumentów, ale nie powinna być traktowana jako pełny zamiennik rzeczywistego udziału różnych grup w procesie deliberacji.

Ta różnica ma kluczowe znaczenie polityczne. Demokratyczna debata to nie tylko algorytm generujący szeroki katalog argumentów; obejmuje ona prawo osób, których dotyczą decyzje, do występowania we własnym imieniu. System może perfekcyjnie symulować argumentację danej grupy społecznej, a mimo to nie zastąpić jej politycznej reprezentacji. Epistemiczna różnorodność i demokratyczna reprezentacja są pojęciami powiązanymi, lecz nierównoważnymi. Hybrydowa Ludzkość potrzebuje zatem nie tylko modeli potrafiących „widzieć wiele perspektyw", lecz instytucji umożliwiających realnym ludziom kwestionowanie sposobu, w jaki ich perspektywy zostały przez system zrekonstruowane.

Jest to epistemiczny odpowiednik contestability omawianej wcześniej. Jeśli AI informuje administrację, że „interesy danej grupy są następujące", członkowie tej grupy muszą zachować prawo do odpowiedzi: „to nie jest adekwatna reprezentacja naszego stanowiska". Jeszcze bardziej złożony problem dotyczy personalizacji epistemicznej. Dzisiejsze systemy informacyjne dostosowują rekomendacje do historii użytkownika, lecz generatywna AI może pójść znacznie dalej: nie tylko decydować, którą wiadomość wyświetlić, ale zmieniać język, argumentację, kolejność informacji, przykłady i stopień szczegółowości w zależności od indywidualnego profilu. Dwie osoby pytające o ten sam problem mogą otrzymać nie tylko różne informacje, ale zupełnie inne ramy poznawcze zagadnienia. Personalizacja sama w sobie może zwiększać zrozumienie – podobnie jak dobry nauczyciel dostosowuje sposób wyjaśnienia do ucznia.

Problem pojawia się, gdy adaptacja jest niejawna, użytkownik nie zna kryteriów jej działania, a funkcją celu systemu nie jest maksymalizacja wiedzy odbiorcy. Wtedy personalizacja może tworzyć generatywne bańki poznawcze znacznie subtelniejsze od klasycznych filter bubbles. Tradycyjna bańka filtrująca ogranicza przede wszystkim ekspozycję; generatywna bańka może natomiast ograniczać interpretację.

Sycophancy i ryzyko potwierdzania błędnych przekonań

System nie musi ukrywać argumentu przeciwnego; może przedstawić go w formie słabszej, mniej przekonującej lub pozbawionej kontekstu. Może ułożyć fakty w kolejności wspierającej wcześniejszą intuicję użytkownika, a kierując się zasadą pomocności i unikania konfliktu – nadmiernie potwierdzać jego perspektywę. W skrajnych przypadkach personalizowany chatbot może uczestniczyć we wzajemnym wzmacnianiu błędnych przekonań, co współczesna literatura analizuje jako szczególną klasę feedback loops. To sprawia, że pojęcie sycophancy ma znaczenie szersze niż zwykły problem jakości odpowiedzi. Model przesadnie dostosowujący się do stanowiska użytkownika może być społecznie przyjemny, ale epistemicznie niebezpieczny. Partner poznawczy powinien czasem potwierdzać, a czasem sprzeciwiać się. Jeśli funkcją systemu staje się maksymalizowanie subiektywnego poczucia satysfakcji, może on przekształcić się z narzędzia poznania w generator uzasadnień dla już posiadanych przekonań. Dojrzały agent epistemiczny powinien więc posiadać nie tylko zdolność generowania odpowiedzi, ale również habit of refusal, o którym Gujral wspomina w kontekście epistemicznej pokory. System powinien umieć powiedzieć: „nie wiem", „dowody są sprzeczne", „twoje założenie nie jest potwierdzone", „nie mam dostępu do źródła pozwalającego to ustalić" albo „istnieje silny kontrargument, którego nie powinieneś pomijać". Odmowa epistemiczna nie jest defektem inteligencji – jest warunkiem jej wiarygodności. Najwyższa kompetencja poznawcza może wymagać kontrolowanej niedomkniętości. Człowiek często interpretuje płynną odpowiedź jako sygnał wiedzy, tymczasem prawdziwie eksperckie stanowisko w wielu dziedzinach brzmi: „wyniki nie pozwalają jeszcze rozstrzygnąć". AI trenowana przede wszystkim do dostarczania użytecznych odpowiedzi może znajdować się pod funkcjonalną presją, aby każdą lukę poznawczą przekształcić w narrację. System epistemicznie dojrzały powinien potrafić utrzymać lukę jako lukę. Jest to bezpośrednio związane z rozróżnieniem truthfulness i helpfulness. Odpowiedź może być pomocna, bo daje użytkownikowi jasny kierunek, lecz prawdziwa struktura wiedzy może pozostać niejasna. W medycynie, ekonomii czy polityce najbardziej uczciwa odpowiedź często wymaga przedstawienia kilku scenariuszy i zakresów niepewności. Projektowanie systemu powinno więc rozróżniać sytuację, w której „użytkownik otrzymał rozwiązanie", od tej, w której „użytkownik został epistemicznie właściwie poinformowany". W tym miejscu można powrócić do źródłowego Explanation Engine.

Wyjaśnialność krytyczna przeciwko syntetycznemu konsensusowi

Gujral przedstawia hybrydę człowieka i AI jako potencjalny „Silnik Wyjaśnień", który ma wykraczać poza prostą predykcję, by zwiększyć ludzką zdolność rozumienia rzeczywistości. Idea ta jest filozoficznie inspirująca, jednak należy odróżnić samo generowanie wyjaśnień od posiadania mocy wyjaśniającej. Płynny tekst odpowiadający na pytanie „dlaczego?" nie staje się automatycznie wyjaśnieniem w sensie naukowym. Prawdziwe wyjaśnienie wymaga bowiem zależności między explanandum a explanans: konkretnego mechanizmu, prawa, struktury przyczynowej lub modelu, który wykaże, dlaczego dane zjawisko wystąpiło. Model językowy potrafi znakomicie reprodukować znane człowiekowi wyjaśnienia, syntetyzować literaturę i proponować hipotezy, ale może również tworzyć historie post hoc, które brzmią przyczynowo, choć rzeczywista zależność nie została wykazana.

Explanation Engine musi zatem opierać się na epistemic provenance, testowalności oraz wyraźnym rozróżnieniu predykcji od przyczynowości. David Deutsch w filozofii nauki kładzie szczególny nacisk na wartość dobrych wyjaśnień, lecz nie oznacza to, że każda narracja wygenerowana przez system posiada explanatory power. Inspiracja Deutschem powinna być tu odczytana jako wezwanie do zachowania możliwości krytycznej rekonstrukcji procesu poznawczego, a nie dowód na to, że odpowiednio rozbudowana AI automatycznie stanie się „silnikiem prawdy". Sam materiał wskazuje, że warunkiem zachowania sprawczości jest transparentność ścieżki decyzyjnej i możliwość krytycznego odrzucenia sugestii maszyny. W tym kontekście wyjaśnialność krytyczna może okazać się ważniejsza niż wyjaśnialność retoryczna. Użytkownik nie potrzebuje jedynie odpowiedzi na pytanie „dlaczego model tak uważa?". Powinien móc zapytać: „co musiałoby być prawdą, aby ten wniosek był fałszywy?", „które założenie ma największy wpływ na wynik?", „które źródła są od siebie zależne?", „jakie dane przemawiają przeciw tej interpretacji?" oraz „czy istnieje model alternatywny równie dobrze wyjaśniający obserwację?". Taki system nie tylko produkuje odpowiedzi, lecz organizuje procedurę falsyfikacji własnych wyników.

Na poziomie społecznym podobną funkcję pełnią instytucje kontroli wiedzy: recenzja naukowa, sądowa kontradyktoryjność, wolne media, audyt i debata publiczna. Nie gwarantują one prawdy, lecz tworzą mechanizmy, dzięki którym twierdzenia napotykają opór. Jeśli AI zastąpi te etapy jednolitą syntezą, może drastycznie zmniejszyć epistemiczne tarcie, które dotąd pełniło funkcję korekcyjną. Hybrydowe społeczeństwo potrzebuje więc instytucjonalnego red teamingu wiedzy. Nie chodzi tu jednak o fetyszyzowanie konfliktu; permanentna polaryzacja niszczy bowiem zdolność do osiągania rzeczywistego porozumienia. Celem nie jest maksymalizacja niezgody, lecz zachowanie

możliwości niezależnej kontroli przed konsensusem. Konsensus jest epistemicznie wartościowy właśnie dlatego, że powstaje po konfrontacji alternatyw, a nie przed nią.

Pozwala to odróżnić konsensus emergentny od konsensusu syntetycznego. Pierwszy powstaje, gdy wiele częściowo niezależnych podmiotów dochodzi do zbliżonych wniosków w drodze badania dowodów i wzajemnej krytyki. Drugi pojawia się, gdy jeden system generuje wypowiedź reprezentującą „to, co wiadomo", a użytkownicy zaczynają ją powtarzać. Choć językowo mogą brzmieć podobnie, ich struktura epistemiczna jest całkowicie inna.

Problem ten pogłębi się, gdy treści generowane przez AI staną się istotną częścią przyszłych danych treningowych i systemy zaczną operować na ideach sformułowanych wcześniej przez inne modele. Nie każdy taki proces prowadzi jednak do „model collapse" w sensie technicznym. Z perspektywy społecznej groźniejsza może być epistemic recursion: sytuacja, w której ludzie i modele cytują oraz reinterpretują treści, których pierwotne źródło staje się nieuchwytne. Choć podobne rekurencje – mity, plotki czy popularne „fakty" – istnieją w kulturze od dawna, skala zjawiska w erze AI nadaje mu nowy wymiar.

Od epistemicznej entropii do infrastruktury autentyczności

AI zwiększa jednak przepustowość generowania treści oraz zdolność do nadawania im stylistycznych pozorów kompetencji. W rezultacie koszt produkcji tekstu, który brzmi przekonująco, dramatycznie maleje. Prowadzi to do asymetrii między produkcją a weryfikacją: wygenerowanie tysiąca wiarygodnie brzmiących twierdzeń jest znacznie tańsze niż ich sprawdzenie. Tę asymetrię można nazwać epistemic denial of service.

Społeczeństwo nie musi zostać przekonane do jednej, konkretnej fałszywej narracji. Wystarczy zalać przestrzeń informacyjną tak wieloma wariantami treści, że zdolność instytucji do ich weryfikacji stanie się niewystarczająca. Szkoda polega wówczas na obniżeniu ogólnego zaufania do całego środowiska informacyjnego. Człowiek, nie wiedząc, czemu wierzyć, może wycofać się do źródeł zgodnych z jego tożsamością grupową. Zagrożeniem dla demokracji może być zatem nie tylko dezinformacja, ale epistemiczna entropia. Jeśli koszt ustalenia prawdy rośnie szybciej niż koszt produkcji przekonującej treści, instytucje wiedzy znajdą się pod coraz większą presją. Dlatego przyszłe systemy potrzebują nie tylko detektorów fałszu, lecz infrastruktury autentyczności, provenance i powiązania twierdzeń z odpowiedzialnymi instytucjami.

Interesującą odpowiedź proponuje rozwijana w 2025 roku koncepcja constitutive knowledge sources. Zamiast oczekiwać, że nieprzejrzysty model sam stanie się ostatecznym źródłem epistemicznego autorytetu, system może odwoływać się do instytucji, które konstytuują określone fakty: sądu w odniesieniu do orzeczenia, organu rejestrowego wobec statusu prawnego, urzędu statystycznego wobec danych oficjalnych czy czasopisma wobec opublikowanego badania. W takim modelu AI pełni funkcję mediatora, podczas gdy autorytet pozostaje zakotwiczony instytucjonalnie. Jest to kluczowy kierunek projektowy.

AI nie musi być epistemicznym suwerenem. Może stanowić warstwę dostępu do systemu społecznej wiedzy, której zadaniem jest odnajdywanie, porządkowanie i wyjaśnianie źródeł, bez zastępowania ich własną syntetyczną narracją tam, gdzie autorytet powinien pozostać związany z konkretną instytucją. Należy jednak pamiętać, że instytucje także mogą się mylić; official source nie jest synonimem prawdy uniwersalnej. Istnieje różnica między źródłem konstytutywnym a opisowym. Urząd może konstytuować fakt zarejestrowania spółki, ale nie konstytuuje prawdy o całej gospodarce. Sąd konstytuuje określony stan prawny, lecz jego uzasadnienie może być później poddane krytyce. AI powinna zatem umieć rozróżniać autorytet instytucjonalny od autorytetu epistemicznego.

Inny problem dotyczy danych treningowych i reprezentacji kultur. LLM uczą się na ogromnych korpusach, jednak dostępność treści nie jest równomierna. Języki dominujące, społeczeństwa o wysokim poziomie cyfryzacji oraz grupy częściej publikujące w internecie są silniej reprezentowane niż kultury ustne, małe wspólnoty językowe czy wiedza lokalna. Model optymalizowany pod kątem statystycznego dopasowania może więc przedstawiać kulturę dominującą jako niewidzialny punkt zerowy, a inne perspektywy traktować jako odstępstwa wymagające doprecyzowania. Homogenizacja nie jest zatem jedynie problemem estetycznym.

Może ona prowadzić do epistemic injustice. Miranda Fricker rozróżnia testimonial injustice, gdy wiarygodność osoby jest systematycznie zaniżana z powodu społecznego uprzedzenia, oraz hermeneutical injustice, gdy pewna grupa dysponuje mniejszym dostępem do wspólnych zasobów pojęciowych pozwalających opisać własne doświadczenia. AI może reprodukować obie te formy: marginalizować określone źródła albo narzucać dominujący aparat pojęciowy do opisu doświadczeń grup, które wytworzyły inne kategorie. W Hybrydowej Ludzkości pojawia się więc potrzeba epistemicznej reprezentacji infrastrukturalnej. Nie wystarczy, aby model był statystycznie „mniej stronniczy". Trzeba pytać, które tradycje wiedzy, języki i sposoby argumentacji pozostają w nim dostępne oraz czy użytkownik ma możliwość wyjścia poza dominującą syntezę. Jednocześnie nie należy romantyzować każdej wiedzy lokalnej ani traktować pluralizmu jako ochrony twierdzeń fałszywych. Tradycja kulturowa może zawierać głęboką wiedzę środowiskową i jednocześnie błędne

przekonania przyrodnicze. Nauka także bywa historycznie obciążona instytucjonalnymi biasami, ale posiada rozwinięte mechanizmy kontroli empirycznej. Zadaniem epistemologii hybrydowej jest łączenie pluralizmu źródeł z asymetrią kryteriów uzasadnienia.

Od roli wyroczni do mapowania krajobrazu wiedzy

W tym miejscu warto zastosować zasadę steelman, która jest kluczowa również dla całego niniejszego artykułu. System powinien prezentować najsilniejszą możliwą wersję znaczącego stanowiska przeciwnego, a nie karykaturę łatwą do odrzucenia – co staje się szczególnie istotne w tematach spornych.

Syntetyczny mediator może podnieść jakość deliberacji, o ile potrafi przedstawić konflikt argumentów z większą uczciwością niż spolaryzowane media. Istnieje jednak ryzyko wytwarzania fałszywej symetrii, gdy stanowiska drastycznie różniące się pod względem dowodowym są prezentowane jako równoważne. Epistemiczny pluralizm wymaga zatem jednoczesnej reprezentacji zarówno różnic w stanowiskach, jak i różnic w ich sile dowodowej. Nie należy usuwać istotnych hipotez mniejszościowych, ale nie wolno też przedstawiać ich jako połowy konsensusu, jeśli popiera je zaledwie niewielka część badań. Jest to zadanie trudne, ponieważ język naturalny zbyt łatwo sprowadza złożony rozkład wiarygodności do formuły binarne "istnieją dwie strony".

Tu właśnie pojawia się problem epistemicznej kartografii. Przyszłe systemy AI powinny być zdolne nie tylko do odpowiadania na pytania, ale przede wszystkim do mapowania struktury wiedzy: głównych teorii, zakresu konsensusu, poziomu niepewności, argumentów mniejszościowych, krytycznych eksperymentów, luk w danych oraz zależności między źródłami. Taka mapa byłaby naukowo cenniejsza niż najbardziej płynna, pojedyncza odpowiedź. Można to porównać do różnicy między przewodnikiem a mapą. Przewodnik mówi: „idź tędy”. Mapa pozwala natomiast zobaczyć teren i dokonać własnego wyboru. Dzisiejsze systemy generatywne są często doskonałymi przewodnikami, jednak hybrydowa sprawczość wymaga, aby w kluczowych kwestiach mogły stać się mapą epistemicznego krajobrazu. Właśnie wtedy AI wzmacnia, zamiast redukować ludzką metakognicję.

Użytkownik nie musi pamiętać wszystkich źródeł, ale powinien otrzymać strukturę pozwalającą zrozumieć, gdzie wiedza jest pewna, a gdzie słaba. W ten sposób cognitive offloading nie prowadzi do epistemicznego podporządkowania – człowiek outsourcinguje proces przeszukiwania ogromnych zbiorów danych, lecz zachowuje kontrolę nad kryteriami uzasadnienia. Na tej podstawie można wyróżnić trzy modele relacji epistemicznej z AI.

W modelu oracle system odpowiada, a człowiek akceptuje. W modelu assistant system dostarcza dowodów i propozycji, podczas gdy człowiek zachowuje funkcję oceny. W modelu epistemic environment AI organizuje całą przestrzeń źródeł, alternatyw i kontrargumentów, umożliwiając człowiekowi swobodne poruszanie się po niej. Pierwszy model jest najwygodniejszy, lecz najbardziej podatny na autorytarną zależność epistemiczną. Trzeci jest najbardziej wymagający projektowo, ale najbliższy ideałowi augmentacji sprawczości. W praktyce wszystkie trzy będą współistnieć.

Przejście od wygody użytkownika ku transparentności epistemicznej

Przy pytaniu o pogodę lub prostą definicję użytkownik potrzebuje oracle. W medycynie, prawie, nauce i polityce wymaga jednak raczej assistant lub epistemic environment. Głębokość uzasadnienia powinna być proporcjonalna do stawki decyzji, podobnie jak wcześniej autonomia agenta była zależna od odwracalności skutku. Epistemologia hybrydowa potrzebuje zatem zasad graduated epistemic transparency. Nie każda odpowiedź wymaga pełnego grafu źródeł; przy decyzjach o wysokiej stawce system powinien jednak ujawniać więcej: źródła, aktualność, niezależność, stopień konsensusu, ograniczenia, alternatywy oraz potencjalny konflikt interesów. Szczególnie istotna będzie informacja o nieobecności wiedzy. Dzisiejsza kultura cyfrowa przyzwyczaiła nas do przekonania, że na każde zapytanie „coś się znajdzie", a generatywna AI dodatkowo eliminuje pustą stronę wyników. Tymczasem w nauce „brak danych" sam w sobie jest informacją. System powinien umieć wskazać, że określone pytanie nie zostało dostatecznie zbadane, zamiast syntetyzować odpowiedź z materiałów jedynie częściowo powiązanych. Ma to fundamentalne znaczenie dla decyzji publicznych: jeśli administracja pyta model o skutki nowej polityki, może otrzymać szczegółową prognozę nawet wtedy, gdy odpowiednie dane historyczne nie istnieją.

Płynność języka potrafi zamaskować epistemic extrapolation. Właściwa odpowiedź powinna zatem oddzielać to, co wynika z danych, od tego, co jest modelem, analogią lub normatywnym założeniem. Zasada ta odpowiada głównemu rygorowi całego artykułu: fakt, model, hipoteza, analogia i norma muszą pozostawać rozróżnialne. W społeczeństwie hybrydowym zadanie to przestaje być wyłącznie obowiązkiem autora naukowego – musi stać się immanentną cechą infrastruktury poznawczej.

W tym miejscu można ponownie odczytać źródłową tezę o „multi-biased future". Jej najbardziej wartościowym rozwinięciem nie jest wielość tendencyjnych modeli, lecz architektura epistemicznej kontradyktoryjności. System może uruchamiać kilka niezależnych sposobów analizy, wykrywać

obszary zgodności i niezgody, a następnie przedstawiać użytkownikowi strukturę sporu. Tak skonstruowana AI nie usuwa konfliktu z wiedzy; ona uczy człowieka, jak ten konflikt wygląda. Taki system może być poznawczo mniej „przyjemny" – jedna pewna odpowiedź jest łatwiejsza do przetworzenia niż pięć wariantów o różnych zakresach niepewności. Epistemicznie odpowiedzialna AI może więc niekiedy zmniejszać wygodę użytkownika, a właśnie dlatego projektowanie prawdy może pozostawać w konflikcie z projektowaniem frictionless experience. Ten konflikt stanowi jeden z najbardziej podstawowych problemów ekonomii produktów AI.

Jeśli użytkownicy preferują jasne i szybkie odpowiedzi, rynek będzie nagradzać systemy generujące silniejsze poczucie pewności. Epistemiczna uczciwość wymaga natomiast częstszych zastrzeżeń, wskazywania alternatyw i przyznawania się do niewiedzy. Bez dodatkowych norm konkurencja o doświadczenie użytkownika może premiować perswazyjną płynność nad intelektualną ostrożnością. Z tego względu polityka AI nie może sprowadzać się jedynie do minimalizowania liczby fałszywych zdań. Potrzebne są normy dotyczące struktury epistemicznego wpływu: wyraźne oznaczanie niepewności, rozróżnienie źródła pierwotnego od wtórnego, możliwość prześledzenia zależności cytowań, ujawnianie istotnych kontrowersji oraz prawo użytkownika do uzyskania alternatywnej interpretacji. Najnowsza literatura regulacyjna wskazuje, że koncentracja wyłącznie na statycznej accuracy może być paradoksalnie niewystarczająca, gdyż szkody generatywnej AI obejmują również subtelną perswazję, iluzję konsensusu i społeczne zubożenie przestrzeni idei. Jest to kluczowy moment dojrzewania governance AI: system można uznać za technicznie „bardzo dokładny", a jednocześnie za epistemicznie niebezpieczny, jeśli jego powszechne użycie zmniejsza różnorodność lub zdolność do niezależnej weryfikacji. Właśnie dlatego wskaźniki jakości przyszłych systemów powinny obejmować nie tylko accuracy, calibration i robustness, ale także source diversity, disagreement preservation, provenance fidelity, uncertainty communication oraz collective diversity impact.

Pluralizm infrastrukturalny jako warunek odporności poznawczej

Jest to trudne metodologicznie, lecz możliwe. Badania nad homogenizacją kreatywną już teraz próbują mierzyć diversity growth rate oraz inne właściwości rozkładu zbiorowych wyników. Analogicznie można badać, czy miliony interakcji z modelem zwiększają, czy raczej zmniejszają zakres publicznie dostępnych argumentów. Nie jest to postulat, aby model celowo generował błędne stanowiska w imię „różnorodności" – epistemiczna różnorodność nie jest bowiem losowością. Dobry system powinien maksymalizować liczbę uzasadnionych hipotez, zachowując

jednocześnie zdolność do ich rankingowania na podstawie dowodów. Najbardziej wartościowy partner poznawczy przypomina zatem dobre seminarium naukowe: nie odpowiada wszystkim jednym głosem.

Prezentuje najsilniejsze argumenty, wskazuje punkty zapalne i proponuje eksperymenty mogące rozstrzygnąć spór. Nie pozwala mniejszościowej teorii zniknąć tylko dlatego, że występuje rzadziej w korpusie danych, lecz nie przedstawia jej również jako równorzędnej z teorią wielokrotnie lepiej potwierdzoną. Epistemiczna różnorodność pełni tu funkcję adaptacyjną. Społeczeństwo dysponujące wieloma niezależnymi modelami świata jest bardziej odporne na sytuację, w której dominujący paradygmat okazuje się błędny. Podobnie ekosystem biologiczny oparty na monokulturze może być niezwykle wydajny w stabilnych warunkach, ale pozostaje szczególnie podatny na nowy patogen. Kognitywna monokultura może przyspieszać koordynację, lecz jednocześnie obniża odporność na błąd systemowy.

To prowadzi do pojęcia epistemic redundancy. W inżynierii dwa identyczne czujniki zwiększają niezawodność w obliczu losowej awarii, ale nie chronią przed wspólnym błędem konstrukcyjnym. Podobnie pięć modeli trenowanych podobnymi metodami na zbliżonych danych może dostarczyć pięciu zgodnych odpowiedzi, nie zapewniając przy tym pięciu niezależnych potwierdzeń. Prawdziwa redundancja wymaga odmiennych metod, danych i założeń. Zasada ta powinna zmienić sposób wykorzystywania multi-agent systems. Zespół agentów nie staje się epistemicznie pluralistyczny samym faktem istnienia wielu instancji. Jeśli wszystkie korzystają z jednego modelu i podobnego promptu, są raczej wielokrotnymi kopiami tego samego eksperta niż panelem niezależnych specjalistów. Liczba agentów nie jest miarą liczby perspektyw.

W systemie Hybrydowej Ludzkości niezbędna będzie kategoria epistemic independence score: wskaźnik stopnia, w jakim poszczególne źródła lub agenci dostarczający zgodnych informacji są rzeczywiście niezależni pod względem danych, metod i mechanizmu generowania. Bez tego syntetyczny konsensus może coraz częściej imitować mądrość tłumu, będąc w rzeczywistości wielokrotnym echem jednego mechanizmu.

Problem ten dotyczy także obywateli. Jeżeli milion ludzi formułuje własne stanowiska polityczne po konsultacji z tym samym systemem, demokratyczny pluralizm liczebny może skrywać korelację upstreamową. Każdy obywatel oddaje osobny głos, lecz część argumentacji została wcześniej ukształtowana przez wspólną infrastrukturę. Nie oznacza to automatycznej manipulacji; gazety, szkoły i kultura od dawna tworzą wspólne ramy poznawcze. AI zwiększa jednak skalę, personalizację i prędkość tego procesu.

Z tego powodu infrastruktura demokratyczna powinna chronić nie tylko wolność wypowiedzi, ale również pluralizm infrastruktury wypowiedzi. Jeśli wszystkie strony politycznego sporu używają jednego systemu do przygotowywania komunikacji, może on stać się ukrytym współautorem języka całej debaty. Wówczas kwestia ustawień modelu zyskuje znaczenie ustrojowe, nawet bez intencji wpływania na wybory. Nie należy jednak wyciągać z tego wniosku, że AI koniecznie osłabi demokrację.

Równie prawdopodobne są zastosowania wzmacniające deliberację: streszczanie projektów ustaw, tłumaczenie języka prawnego, udostępnianie kontrargumentów, symulowanie konsekwencji, wykrywanie niespójności i ułatwianie obywatelom dostępu do wiedzy eksperckiej. Generatywna AI może obniżyć bariery wejścia do debaty publicznej tak radykalnie, jak niegdyś druk i internet.

Epistemiczne zarządzanie zamiast ślepego zaufania

Pytanie brzmi, czy system będzie działać jako wielogłosowy tłumacz wspólnej wiedzy, czy raczej jako jeden domyślny narrator rzeczywistości. W tym miejscu dochodzimy do centralnej zasady epistemologii Hybrydowej Ludzkości: AI powinna zwiększać zdolność społeczeństwa do wytwarzania uzasadnionej niezgody i uzasadnionego konsensusu, a nie maksymalizować sam konsensus. Celem nie jest bowiem zgodność sama w sobie, lecz proces, w którym zgoda – gdy już się pojawia – zasługuje na zaufanie.

To rozróżnienie jest kluczowe w kontekście wizji Supersociety, gdzie zaufanie ma budować przejrzysty przepływ informacji wewnątrz zintegrowanego systemu człowiek-maszyna. Sama transparentność nie wystarcza jednak do zbudowania zaufania; system może być perfekcyjnie przejrzysty i jednocześnie błędny. Zaufanie społeczne wymaga bowiem niezależnej kontroli, pluralizmu, odpowiedzialności, prawa sprzeciwu oraz możliwości wyjścia z dominującej infrastruktury. W społeczeństwie hybrydowym informacja stanie się prawdopodobnie coraz łatwiej dostępna, za to jej uzasadnienie – coraz cenniejsze. Problem niedoboru przesuwa się: o ile XX wiek charakteryzował się wysokim kosztem dostępu do informacji, o tyle XXI wiek cechuje nadmiar danych przy jednoczesnym braku mechanizmów pozwalających ustalić, którym z nich warto zaufać. AI pogłębia oba te zjawiska: zwiększa dostępność treści, ale jednocześnie mnoży ilość materiałów syntetycznych.

Można zatem przewidywać wzrost wartości epistemicznych instytucji kotwiczących: laboratoriów, bibliotek, archiwów, rejestrów publicznych, czasopism naukowych, standardów provenance, systemów certyfikacji oraz niezależnych audytów. Nie chodzi o to, że instytucje te są nieomylne, lecz o to, że umożliwiają one rekonstrukcję odpowiedzialności za konkretne twierdzenie. Właśnie ta

odpowiedzialność jest kategorią, której brakuje anonimowej generacji treści. Człowiek podpisujący publikację naukową lub opinię ekspercką ryzykuje reputacją i wystawia się na krytykę; LLM nie ponosi analogicznego kosztu społecznego. Dostawca systemu może natomiast ponosić odpowiedzialność organizacyjną, dlatego projekt epistemiczny AI powinien budować łańcuch accountable provenance – od źródła, poprzez system, aż do użytkownika. W ten sposób analiza odpowiedzialności moralnej łączy się z epistemologią.

Jeśli system generuje błąd, nie wystarczy stwierdzenie, że „model się pomylił". Należy ustalić, czy błędne było samo źródło, czy mechanizm retrieval wybrał niewłaściwy materiał, czy model dokonał nieuprawnionej syntezy, czy może użytkownik źle zinterpretował poziom pewności, a interfejs zamaskował istniejącą niezgodność. Odpowiedzialność epistemiczna jest rozproszona, lecz nie powinno to prowadzić do jej rozmycia. Podobnie jak w moral crumple zone, konieczne są wcześniej zdefiniowane role: dostawca odpowiada za określone właściwości systemu, wydawca za źródło, użytkownik profesjonalny za zakres zastosowania, organizacja za procedurę, a instytucja publiczna za możliwość odwołania. Wiedza hybrydowa wymaga zatem epistemic governance, a nie tylko sprawnych modeli.

W dojrzałym systemie można sformułować zasadę czterech warstw: źródło, synteza, niepewność i odpowiedzialność. Użytkownik powinien mieć możliwość ustalenia, skąd pochodzi istotne twierdzenie, jak zostało przekształcone, jaki jest poziom epistemicznej pewności i kto odpowiada za użycie informacji w danym kontekście. Taka infrastruktura mogłaby paradoksalnie zwiększyć jakość epistemiczną względem dzisiejszego internetu. Człowiek samodzielnie wyszukujący informacje często polega na pierwszych wynikach, pomija artykuły naukowe i nie potrafi ocenić metodologii. AI mogłaby natomiast automatycznie wykrywać retrakcje, wskazywać konflikty interesów, porównywać metaanalizy i informować o jakości projektu badawczego. Technologia nie musi więc prowadzić do epistemicznego outsourcingu; może stać się metapoznawczym egzoszkieletem społeczeństwa.

Warunkiem jest jednak właściwa funkcja celu. System projektowany pod kątem szybkości i satysfakcji użytkownika będzie inny od systemu optymalizowanego pod kątem różnorodności epistemicznej i odporności na błędny konsensus. To prowadzi z powrotem do kwestii ekonomii politycznej: kto zapłaci za jakość epistemiczną, której użytkownik nie zawsze potrafi dostrzec? Rynek może bowiem nie wynagradzać wszystkich dóbr epistemicznych, gdyż utrzymywanie alternatywnych analiz, provenance i niezależnej kontroli znacząco zwiększa koszty.

Policentryzm i sprawczość jako tarcza przed homogenizacją wiedzy

Korzyści z różnorodności mają często charakter zbiorowy. Jednostka może racjonalnie wybierać najwygodniejszą odpowiedź, choć społeczeństwo traci przez to na pluralizmie. To klasyczna analogia do problemu dóbr publicznych; epistemiczna różnorodność powinna zatem być traktowana jako dobro wspólne. Podobnie jak finansuje się biblioteki, naukę podstawową czy archiwa, państwo i społeczeństwo mogą mieć interes w utrzymywaniu infrastruktury AI, która zachowuje dostęp do zróżnicowanych źródeł i modeli – nawet jeśli nie jest to najbardziej dochodowa architektura komercyjna. Przyszłe biblioteki mogłyby zatem wyjść poza rolę magazynów tekstów i stać się publicznymi mediatorami AI, pozwalającymi obywatelowi analizować wiedzę bez konieczności podporządkowania się komercyjnemu profilowaniu. Uniwersytety mogą utrzymywać niezależne modele badawcze, media publiczne rozwijać narzędzia provenance, a społeczeństwo obywatelskie budować systemy red-teamingu i audytu.

Właśnie policentryzm może okazać się najlepszą odpowiedzią na epistemiczne ryzyko AI. Nie istnieje jedna instytucja, której można powierzyć kontrolę nad prawdą bez ryzyka nadużyć. Państwo, rynek, nauka i społeczeństwo obywatelskie generują odmienne typy błędów, więc wzajemne ograniczanie się centrów poznawczych stanowi mechanizm bezpieczeństwa, analogiczny do pluralizmu modeli. W takich warunkach najważniejszym zadaniem edukacji przestaje być pamięciowe konkurowanie z AI, a staje się rozwój epistemic agency. Obywatel powinien wiedzieć, kiedy niezbędne jest źródło pierwotne, czym różni się korelacja od przyczynowości i jak rozumieć niepewność. Musi rozumieć, dlaczego konsensus wielu zależnych źródeł nie jest prawdziwym konsensusem oraz kiedy płynny tekst nie stanowi jeszcze dowodu.

AI literacy nie powinna być zatem kursem obsługi promptów, lecz elementem współczesnej teorii obywatelstwa. Człowiek żyjący w środowisku syntetycznych odpowiedzi musi potrafić zarządzać nie tylko informacją, ale i relacją z epistemicznym pośrednikiem. Paradoksalnie im lepsze staną się modele, tym trudniejsza może być ta umiejętność. Oczywiste błędy mobilizują sceptycyzm; system poprawny w 99 przypadkach buduje jednak silniejsze zaufanie, co sprawia, że setny błąd może przejść niezauważony. Wysoka jakość ogólna nie eliminuje więc potrzeby krytycznej kompetencji – zmienia jedynie jej charakter i częstotliwość występowania. Można to porównać do lotnictwa.

Wysoka niezawodność automatyki zwiększa bezpieczeństwo, lecz jednocześnie sprawia, że człowiek rzadziej ćwiczy interwencję w sytuacjach nietypowych. W epistemologii AI analogicznym problemem będzie rare catastrophic belief error: rzadki, lecz niezwykle przekonujący błąd w domenie wysokiej stawki. Dlatego epistemiczna odporność wymaga utrzymywania kompetencji

niezależnych nawet wtedy, gdy są one wykorzystywane rzadko. Lekarz powinien potrafić zakwestionować system diagnostyczny, naukowiec rozpoznać błędną syntezę literatury, prawnik przeczytać orzeczenie pierwotne, a obywatel odróżnić źródło od komentarza. To epistemiczny odpowiednik zasady retained human capability.

Pluralizm perspektyw jako fundament wolności poznawczej

Największym zagrożeniem nie jest świat, w którym AI generuje wyłącznie fałsze – taki system szybko straciłby swą użyteczność. Znacznie bardziej problematyczna i subtelna jest wizja świata, w którym AI jest zazwyczaj trafna, niezwykle wygodna, stylistycznie przekonująca i na tyle jednolita, że stopniowo przekształca społeczne kryteria wiedzy. W takim scenariuszu centralna cenzura nie byłaby potrzebna do ograniczenia pluralizmu. Wystarczy, że większość ludzi będzie korzystać z podobnych systemów, które preferują statystycznie dominujące reprezentacje. Kultura może ulec homogenizacji poprzez optymalizację pod kątem przeciętnej użyteczności, a nie w wyniku politycznego nakazu.

W tym punkcie analiza osiąga swój najistotniejszy wymiar epistemologiczny. Teza, że AI zaczyna kształtować „inputs" ludzkiego umysłu, powinna być traktowana jako realny problem architektury społecznej wiedzy. Jeśli pośrednik selekcjonuje dane wejściowe, dysponuje władzą uprzednią wobec perswazji. Nie musi przekonywać człowieka do sformułowanego już poglądu – może współtworzyć przestrzeń, z której ten pogląd dopiero się wyłoni. Wobec tego wolność epistemiczna nie może ograniczać się jedynie do prawa wypowiedzi; musi obejmować dostęp do pluralistycznego środowiska powstawania przekonań. Jest to rozwinięcie koncepcji cognitive liberty: wolność poznawcza obejmuje nie tylko brak ingerencji w mózg, ale także niedominacyjny dostęp do infrastruktury selekcjonującej i interpretującej informacje. Nie oznacza to obowiązku wystawiania każdego człowieka na każdy możliwy pogląd, co byłoby poznawczo niemożliwe, gdyż każda kognicja wymaga filtrów. Problemem jest kontrola nad filtrem, świadomość jego istnienia oraz możliwość jego zmiany. Ludzki mózg, kultura i AI są pod tym względem podobne: wszystkie selekcjonują. Różnica polega na tym, że filtr AI jest projektowalny, aktualizowany masowo i działa jednocześnie dla milionów osób, co nadaje mu wyjątkowe znaczenie polityczne.

Dojrzałe społeczeństwo hybrydowe nie powinno zatem dążyć do świata „bez biasów", lecz do świata widzialnych założeń, konkurujących modeli, sprawdzalnych źródeł i możliwości korekty. To właśnie jest naukowo sensowna wersja multi-biased future: nie pluralizm błędów, lecz pluralizm kontrolowalnych perspektyw. Z tej perspektywy prawda nie staje się relatywna; wręcz przeciwnie –

wymaga coraz bardziej rozbudowanej infrastruktury, ponieważ droga między człowiekiem a dowodem zostaje wydłużona o kolejne warstwy modeli. Im więcej pośredników, tym ważniejsze stają się provenance, niezależność źródeł i możliwość powrotu do materiału pierwotnego. Hybrydowa epistemologia powinna zatem chronić prawo do zejścia warstwę niżej: użytkownik otrzymujący syntezę powinien móc przejść do źródła, z poziomu źródła – do danych, ze statystyki – do metodologii, a z modelu – do założeń.

Choć nie każdy będzie korzystał z tej możliwości, samo jej istnienie ogranicza arbitralną władzę pośrednika. Jest to epistemiczny odpowiednik prawa do wyjścia z infrastruktury. Exit from synthesis oznacza możliwość opuszczenia gotowej narracji i ponownego wejścia w materiał, z którego została zbudowana. Bez tego AI staje się poznawczym terminalem końcowym. Z tym prawem może pozostać narzędziem orientacji. Można zatem sformułować minimalne zasady epistemologii społeczeństwa hybrydowego: AI powinna jasno oddzielać fakt od hipotezy i normy, ujawniać istotną niepewność, zachowywać znaczącą niezgodność źródeł, umożliwiać rekonstrukcję provenance, rozróżniać źródła zależne od niezależnego potwierdzenia, unikać przedstawiania syntetycznej płynności jako dowodu konsensusu, wspierać pluralizm modeli oraz umożliwiać człowiekowi niezależne zakwestionowanie odpowiedzi. Kluczowym miernikiem nie będzie pytanie „czy użytkownik dostał odpowiedź?", lecz czy po interakcji posiada lepiej skalibrowane przekonania i precyzyjniejszą mapę tego, co wciąż pozostaje nieznane. Jest to wymaganie znacznie wyższe niż conversational helpfulness. Właśnie w tym miejscu Hybrydowa Ludzkość może osiągnąć swoją najbardziej obiecującą postać.

AI nie musiałaby zastępować ludzkiej epistemologii; mogłaby uczynić widoczną strukturę wiedzy, której pojedynczy człowiek nie jest w stanie ogarnąć: miliony badań, sprzeczne dane, różne tradycje i historię zmieniającego się konsensusu. Odpowiednio zaprojektowana hybryda zwiększyłaby nie tylko ilość dostępnej informacji, ale i zdolność społeczeństwa do rozumienia własnej niewiedzy. Taka perspektywa prowadzi jednak bezpośrednio do najbardziej spekulatywnej części wywodu. Jeśli wiele osób, agentów, systemów pamięci i instytucji zostanie połączonych w sieć zdolną do utrzymywania wspólnego modelu wiedzy, pojawia się pytanie o nowy poziom organizacji poznawczej. Źródło definiuje go poprzez kategorie Metamind, Multi-Substrate Person, Supersociety i Explanation Engine, sugerując przejście od indywidualnych umysłów ku wielopodłożowym systemom poznawczym.

Pojęcia te nie są opisem faktów ani nieuchronną prognozą technologiczną. Mogą jednak służyć jako rygorystycznie oznaczona futurologia systemowa. Pytanie nie brzmi, czy ludzkość stworzy mistyczny „superumysł", lecz czy integracja ludzi, agentów, pamięci i instytucji może wytworzyć nowe własności funkcjonalne na poziomie całych społeczeństw – zdolność wyjaśniania, planowania

i koordynacji, której nie posiada żaden pojedynczy element. Dopiero na tym poziomie można zapytać, czy po AGI następuje kolejny rodzaj maszyny, czy raczej zmiana skali jednostki inteligencji. Kolejna część musi zatem przejść od epistemologii hybrydowego społeczeństwa do koncepcji Metamind i Supersociety, zachowując rygorystyczną granicę między teorią systemów złożonych, realnie obserwowalną inteligencją zbiorową a futurologiczną wizją wielosubstratowego podmiotu.

> Czy w świecie, w którym AI potrafi perfekcyjnie symulować każdą możliwą perspektywę, prawdziwy pluralizm stanie się jedynie estetycznym dodatkiem do zunifikowanego algorytmu? Być może największym wyzwaniem przyszłości nie będzie walka z fałszem, lecz opór przeciwko zbyt wygodnej prawdzie. Ostatecznie to właśnie w szczelinach między sprzecznymi opiniami i w dyskomforcie niewiedzy kryje się przestrzeń, w której wciąż możemy nazywać siebie myślącymi podmiotami.

Inspiracje

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

Rana Gujral (ur. 18 czerwca 1976 r. w Pune w Indiach) jest indyjsko-amerykańskim przedsiębiorcą technologicznym, dyrektorem i inwestorem, specjalizującym się w kognitywnej sztucznej inteligencji, sztucznej inteligencji emocjonalnej i uczeniu maszynowym. Wcześniej pełnił funkcję dyrektora generalnego Behavioral Signals, gdzie kierował rozwojem systemów przetwarzania mowy i sygnałów oraz systemów rozpoznawania emocji i zachowań opartych na głosie. Wcześniej Gujral założył TiZE, firmę zajmującą się oprogramowaniem do uczenia maszynowego dla przedsiębiorstw, przejętą przez Alchemy, i piastował stanowiska kierownicze w takich firmach jak Logitech i Cricut. Uznany za wpływowy głos w dziedzinie nowych technologii i sztucznej inteligencji ogólnej (AGI), Gujral opowiada się za zrozumieniem hybrydowego poznania człowieka i sztucznej inteligencji oraz wprowadził koncepcję sztucznego doświadczenia ogólnego (AGE). Często publikuje swoje spostrzeżenia na temat technologii, przywództwa i biznesu w ważnych publikacjach branżowych, takich jak Forbes, Inc. i TechCrunch.

Rana Gujral *(born June 18, 1976, in Pune, India) is an Indian-American technology entrepreneur, executive, and investor specializing in cognitive artificial intelligence, emotion AI, and machine learning. He previously served as the Chief Executive Officer of Behavioral Signals, where he led the development of speech-signal processing and voice-based emotion and behavioral recognition systems. Prior to that, Gujral founded TiZE, an enterprise machine learning software company acquired by Alchemy, and held executive leadership roles at companies including Logitech and Cricut. Recognized as an influential voice in emerging technologies and Artificial General Intelligence (AGI), Gujral advocates for understanding hybrid human-AI cognition and introduced the concept of Artificial General Experience (AGE). He frequently contributes insights on technology, leadership, and business to major industry publications such as Forbes, Inc., and TechCrunch.*

Behavioral Signals

Literatura uzupełniająca

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Reimagining Truth and Trust"

Donghee Shin

2026 | Algorithmic Fact-Verification | DOI: 10.1093/9780197849644.003.0015

Źródło

[2] "Beyond Christianity: A Critique of the Rational Choice Theory of Religion from a Weberian and Comparative Religions Perspective"

Stephen Sharot

2002 | Sociology of Religion | DOI: 10.2307/3712301

Źródło

[3] "Sociology of Religion: Theoretical and Comparative Perspectives"

Lutz F. Kaelber, Malcolm B. Hamilton, Kevin J. Christiano, William H. Swatos, Peter Kivisto, Stephen Sharot

2002 | Teaching Sociology | DOI: 10.2307/3211507

Źródło

[4] "Comparative Perspectives on Judaisms and Jewish Identities"

Stephen Sharot

2010 | Project Muse (Johns Hopkins University)

Źródło

[5] "A critical commentary on gans' "symbolic ethnicity and symbolic religiosity" and other formulations of ethnicity and religion regarding American Jews"

Stephen Sharot

1997 | Contemporary Jewry | DOI: 10.1007/bf02965478

Źródło

[6] "Beyond the Quantum Horizon Once viewed as imposing absolute limits on knowledge and technology, quantum theory is now expanding the power of computers and the vistas of the mind"

D. Deutsch, A. Ekert

2012

Źródło

Artykuły naukowe autora:

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch02

Źródło

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch03

Źródło

[9] "The AI Instinct"

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.fmatter

Źródło

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

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch04

Źródło

[11] "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: fe84e436-cc35-49c3-bb94-c6009190e228