

Architektura aksjologii i płynność inteligencji: stałe kryteria w świecie nieprzewidywalnych systemów poznawczych w świetle koncepcji Douglasa Hofstadtera.



AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Tekst analizuje koncepcję architektury aksjologii jako zestawu trwałych kryteriów normatywnych, które pełnią rolę bezpiecznika w obliczu dynamicznego rozwoju sztucznej inteligencji. Autor argumentuje, że ze względu na nieprzewidywalność przyszłych technologii, etyka AI nie może opierać się na konkretnych rozwiązaniach technicznych, lecz na stabilnych inwariantach, takich jak prawda, odpowiedzialność i autonomia człowieka. Odwołując się do teorii analogii Hofstadtera oraz międzynarodowych ram (UNESCO, OECD), artykuł podkreśla konieczność zachowania rozróżnienia między formą a treścią w systemach generatywnych. Celem jest stworzenie struktury epistemicznej, która chroni podmiotowość człowieka i zapobiega zastąpieniu ludzkiej sprawczości niewidzialnym sterowaniem algorytmicznym.

The text analyzes the concept of the architecture of axiology as a set of permanent normative criteria that act as a safeguard in the face of the dynamic development of artificial intelligence. The author argues that due to the unpredictability of future technologies, AI ethics cannot rely on specific technical solutions, but on stable invariants such as truth, responsibility, and human autonomy. Drawing on Hofstadter's analogy theory and international frameworks (UNESCO, OECD), the article emphasizes the need to maintain the distinction between form and content in generative systems. The goal is to create an epistemic structure that protects human subjectivity and prevents the replacement of human agency with invisible algorithmic control.

Artykuł stawia tezę, że w obliczu radykalnej nieprzewidywalności rozwoju sztucznej inteligencji, jedynym skutecznym bezpiecznikiem jest ustanowienie trwałej architektury aksjologicznej. Autor argumentuje, że skoro nie potrafimy przewidzieć konkretnych form przyszłych systemów poznawczych (ich architektur czy kompetencji), musimy przestać budować normy wokół technologii, a zacząć opierać je na niezmiennych kryteriach funkcjonalnych: prawdzie, odpowiedzialności, autonomii, godności i sprawiedliwości. Tekst proponuje przejście od naiwnego prognozowania daty nadejścia AGI ku „normatywnej odporności”, gdzie systemy oceniamy nie przez pryzmat ich mocy obliczeniowej, lecz wpływu na podmiotowość człowieka. Inteligencja jest tu definiowana nie jako binarna cecha czy suma faktów, lecz jako dynamiczna zdolność do reorganizacji reprezentacji i zarządzania napięciem między stabilnością a zmianą. W tym ujęciu etyka przestaje być zewnętrznym dodatkiem do gotowego produktu, a staje się warstwą konstytutywną projektu, chroniącą człowieka przed redukcją do poziomu wydajności poznawczej w świecie coraz bardziej płynnych systemów.

SŁOWA KLUCZOWE

architektura aksjologii

płynność inteligencji

Douglas Hofstadter

systemy poznawcze

inwarianty etyczne

autonomia człowieka

epistemiczna symetria

metapoznanie funkcjonalne

normatywna odporność

human agency

efekt ELIZA

poszarpana inteligencja

struktura FARG

rozproszone sprawstwo przyczynowe

Trwałe kryteria aksjologiczne jako bezpiecznik nieprzewidywalności

Architektura aksjologii pozostaje: stałe kryteria w świecie płynnych systemów poznawczych

Formuła „Architektura aksjologii pozostaje” stanowi jedną z nadrzędnych tez całego wywodu autora, obok równie istotnego stwierdzenia, że przyszłość wymyka się predykcji. Zestawienie tych dwóch myśli prowadzi do problemu znacznie głębszego niż standardowa dyskusja o „etyce AI”. Jeżeli nie potrafimy wiarygodnie przewidzieć przyszłych form sztucznej inteligencji – jej architektur, zakresu autonomii, sposobów uczenia czy miejsca w organizacji społecznej – nie

możemy budować porządku normatywnego wyłącznie wokół konkretnych rozwiązań technicznych. Potrzebujemy kryteriów trwalszych niż chwilowa postać technologii. Architektura aksjologii jest właśnie takim poziomem abstrakcji: nie służy szczegółowemu przewidywaniu przyszłości, lecz określa warunki, w których różne scenariusze przyszłe będzie można oceniać.

Jest to strukturalnie zbliżone do problemu inwariancji, który towarzyszył programowi FARG. W systemie Copycat głęboka analogia nie wymaga zachowania wszystkich własności porównywanych sytuacji; przeciwnie, wiele cech powierzchniowych musi ulec przesunięciu, aby mogła zostać zachowana relacja ważniejsza. Autor książki ujmuje tę zasadę w twierdzeniu, że pojęcia głębokie powinny być bardziej stabilne podczas dobrej analogii niż pojęcia powierzchniowe. W perspektywie aksjologicznej można sformułować analogiczną intuicję: rozwój technologii wymaga ogromnej elastyczności na poziomie narzędzi, procedur i organizacji pracy, lecz elastyczność ta nie powinna automatycznie obejmować wartości, według których oceniamy skutki działania systemu.

Nie oznacza to ustanowienia niezmiennego katalogu norm istniejących poza historią. Wartości również podlegają interpretacjom, konfliktom i społecznej reinterpretacji. Godność, autonomia, odpowiedzialność czy sprawiedliwość nie są prostymi zmiennymi technicznymi, a różne tradycje filozoficzne nadają im odmienne znaczenia. Teza o trwałości architektury aksjologicznej jest subtelniejsza: pomimo przemian technologicznych pewne typy pytań normatywnych pozostają nieusuwalne.

Niezależnie od tego, czy system będzie transformerem, architekturą hybrydową, agentem ucieleśnionym czy strukturą jeszcze nieznaną, nadal będziemy musieli pytać: czy jego zastosowanie respektuje osobę? Kto odpowiada za konsekwencje działania? Czy istnieje możliwość sprzeciwu? Czy informacja jest przedstawiana zgodnie z prawdą? Komu przypadają korzyści i ryzyka oraz czy autonomia człowieka nie została zastąpiona niewidzialnym sterowaniem? Współczesne międzynarodowe ramy etyki i zarządzania AI wskazują na podobną trwałość pewnych kryteriów. Rekomendacja UNESCO dotycząca etyki sztucznej inteligencji stawia w centrum godność i prawa człowieka, wymieniając obok nich między innymi przejrzystość, sprawiedliwość oraz ludzką kontrolę. Zaktualizowane w 2024 roku zasady OECD obejmują autonomię jednostki, prywatność, niedyskryminację, przejrzystość, bezpieczeństwo i odpowiedzialność. Konwencja ramowa Rady Europy o AI również organizuje ocenę systemów wokół godności i autonomii, transparentności i nadzoru, odpowiedzialności, równości, prywatności, niezawodności oraz bezpiecznej innowacji. Zbieżność ta nie dowodzi filozoficznej uniwersalności każdego z tych pojęć, lecz pokazuje, że wraz ze zmianą technologii dyskurs normatywny nie porzucił klasycznych pytań o człowieka jako podmiot, lecz próbuje je przełożyć na nowe warunki społeczno-techniczne.

Prawda, odpowiedzialność i autonomia jako niezmiennie filary oceny systemów

Pierwszym takim inwariantem jest **prawda**, rozumiana nie jako naiwne założenie, że każdy komunikat ma prostą wartość binarną, lecz jako zobowiązanie do zachowania rozróżnienia pomiędzy faktem, hipotezą, interpretacją, symulacją a fikcją. Sam prompt metodologiczny tej pracy nakazuje właśnie takie rozgraniczenie: faktów nie należy mieszać z hipotezami, metafor z mechanizmami, korelacji z przyczynowością ani opinii z wynikami badań. Reguła ta ma znaczenie wykraczające poza samą redakcję artykułu.

W świecie systemów generatywnych płynność języka może maskować różnicę pomiędzy wiedzą a prawdopodobnym uzupełnieniem treści. Im doskonalsza staje się stylistyczna wiarygodność systemu, tym ważniejsze jest utrzymanie architektury epistemicznej, która nie utożsamia przekonującej formy z prawdziwością treści. Prawda jest tu przeciwieństwem nie tyle błędu, co **nierozpoznanej niepewności**. Błąd może wystąpić w każdym systemie poznawczym, także ludzkim. Problem normatywny pojawia się wtedy, gdy architektura komunikacji zaciera różnicę pomiędzy pewnością a domysłem lub gdy użytkownikowi uniemożliwia się rozpoznanie źródeł i granic twierdzenia. Dlatego przejrzystość oraz możliwość zakwestionowania rezultatu pojawiają się w aktualnych ramach zarządzania AI jako wymagania związane z odpowiedzialnością. OECD podkreśla potrzebę przekazywania informacji o możliwościach i ograniczeniach systemów oraz, tam gdzie jest to użyteczne, danych umożliwiających osobom dotkniętym decyzją zrozumienie i zakwestionowanie wyniku.

W tym miejscu prawda spotyka się z Hofstadterowską nieufnością wobec wyniku pozbawionego procesu. Jeżeli system podaje odpowiedź, której nie potrafimy powiązać z wiarygodnym mechanizmem, źródłem lub kontrolowalną procedurą, problemem nie musi być fałszywość tej odpowiedzi. Może ona być poprawna, ale epistemicznie krucha. Nauka odróżnia trafienie od uzasadnienia właśnie dlatego, że poznanie wymaga czegoś więcej niż tylko zgodności pojedynczego wyniku z rzeczywistością. Właściwa architektura aksjologiczna powinna więc chronić nie tylko prawdziwość końcowego komunikatu, ale warunki, w których użytkownik może ocenić jego status poznawczy.

Drugim inwariantem jest **odpowiedzialność**. Autor książki wskazuje, że sprawstwo przyczynowe w systemach AI jest rozproszone pomiędzy projektantów, dane, modele, użytkowników, narzędzia i instytucje. Ta dystrybucja nie powinna jednak prowadzić do normatywnego rozpuszczenia odpowiedzialności. Im bardziej złożony staje się łańcuch przyczynowy, tym bardziej potrzebna jest

jasna odpowiedź na pytanie, kto posiadał władzę nad decyzją o zastosowaniu systemu, kto określił jego zakres, kto powinien nadzorować rezultat i kto ma zdolność naprawienia szkody.

Właśnie tutaj należy odróżnić wyjaśnienie od usprawiedliwienia. Stwierdzenie, że "model tak zdecydował", może być elementem wyjaśnienia przyczynowego, lecz nie wystarcza jako usprawiedliwienie instytucjonalne. System techniczny nie powinien stawać się retoryczną zasłoną pozwalającą ludziom i organizacjom uciekać od odpowiedzialności za warunki jego użycia. UNESCO ujmuje tę zasadę bardzo wyraźnie, wskazując, że zastosowanie AI nie powinno wypierać ostatecznej ludzkiej odpowiedzialności i rozliczalności.

Trzecim kryterium jest **autonomia człowieka**. W tradycji kantowskiej autonomia oznacza zdolność podmiotu do kierowania się racjami, które może uznać za własne, zamiast pozostawiania wyłącznie przedmiotem zewnętrznego sterowania. W teorii decyzji i psychologii kwestie te są definiowane nieco inaczej, jednak wspólny rdzeń pozostaje rozpoznawalny: człowiek powinien posiadać rzeczywistą przestrzeń wyboru oraz rozumieć w dostatecznym stopniu warunki, w których dokonuje decyzji. Technologia AI może wspierać tę zdolność, rozszerzając dostęp do informacji, argumentacji i narzędzi poznawczych, lecz może ją również ograniczać, jeżeli system staje się niewidocznym architektem preferencji, selekcji informacji i zakresu dostępnych opcji.

Problem autonomii nie sprowadza się więc do banalnej formuły "człowiek zawsze naciska ostatni przycisk". Formalna obecność człowieka w pętli ma niewielką wartość, jeżeli osoba ta nie dysponuje czasem, informacją, kompetencją albo realną możliwością zakwestionowania wyniku. Nadzór może istnieć nominalnie, a nie funkcjonalnie. Z tego powodu współczesne wytyczne OECD mówią nie tylko o ludzkim nadzorze, ale o zachowaniu **human agency**, czyli rzeczywistej zdolności człowieka do działania i kontroli. Kwestia ta ma interesujący odpowiednik w architekturze poznawczej FARG.

Płynny system nie jest sterowany jednym centralnym modułem, lecz nie oznacza to, że zostaje on pozbawiony organizacji. Lokalna autonomia kodletów funkcjonuje w ramach struktury regulującej ich zasoby, priorytety i wpływ na Workspace. Podobnie społeczeństwo korzystające z autonomicznych systemów nie musi wymagać mikrozarządzania każdym krokiem maszyny, ale musi zachować kontrolę na poziomie reguł, granic i możliwości interwencji. Autonomia komponentu nie może być utożsamiana z brakiem nadrzędnej architektury odpowiedzialności.

Godność jako bariera przed redukcją człowieka do wydajności

Czwartym inwariantem jest **godność**. To pojęcie filozoficznie złożone, gdyż bywa traktowane zarówno jako fundament praw człowieka, jak i ogólne określenie szczególnego statusu osoby. W kontekście AI jego wartość polega na tym, że nie pozwala zredukować człowieka do wydajności poznawczej. Gdyby bowiem wartość ludzka zależała od przewagi w liczeniu, pamięci, diagnozowaniu czy pisaniu, rozwój systemów przekraczających nasze możliwości w kolejnych dziedzinach nieuchronnie prowadziłby do normatywnej degradacji osoby. Koncepcja godności przerywa ten związek: człowiek nie jest podmiotem praw dlatego, że zawsze rozwiązuje zadanie lepiej niż maszyna. Jest to kluczowy punkt debaty o inteligencji. Test Turinga, benchmarki czy pomiary rozumowania dotyczą kompetencji, a ta nie stanowi miary wartości moralnej.

Osoba chora, dziecko, człowiek z poważnymi ograniczeniami poznawczymi czy senior nie posiadają mniejszej godności tylko dlatego, że system komputerowy przewyższa ich w tysiącach zadań. Podobnie wzrost inteligencji maszynowej nie może sam w sobie przekształcić aksjologii osoby w ranking sprawności obliczeniowej. W tym miejscu architektura aksjologiczna chroni przed jednym z najgroźniejszych błędów kategorii: przejściem od stwierdzenia, że „system jest bardziej kompetentny”, do wniosku, iż „system powinien być bardziej uprawniony do decydowania o człowieku”. Kompetencja może uzasadniać delegację konkretnego zadania technicznego, lecz nie ustanawia automatycznie legitymacji moralnej ani społecznej.

Normatywna odporność jako bezpiecznik nieprzewidywalności

System może trafniej prognozować, analizować lub optymalizować, a mimo to decyzja o tym, jakie cele mają być realizowane i jakie koszty są dopuszczalne, pozostaje pytaniem normatywnym. Piątym elementem jest sprawiedliwość, rozumiana nie jako jedna formuła dystrybucyjna, lecz jako pytanie o rozkład korzyści, kosztów, ryzyka oraz możliwości odwołania.

AI może zwiększać wydajność społeczną, jednak efekty technologiczne nie rozkładają się automatycznie równomiernie. Jedna grupa może czerpać korzyści ekonomiczne, podczas gdy inna ponosi ryzyko błędów lub utraty kontroli. Algorytm bywa statystycznie trafny w ujęciu średnim, a jednocześnie znacznie mniej skuteczny dla określonej podpopulacji. Dlatego współczesne ramy etyczne wiążą sztuczną inteligencję z zasadami równości i niedyskryminacji. Z perspektywy teorii

systemów oznacza to, że nie wystarczy optymalizować funkcji celu agregującej średnią korzyść, gdyż taka agregacja może ukrywać lokalne szkody.

Jest to analogiczne do problemu wysokiego wyniku benchmarkowego przy "poszarpanym" profilu kompetencji: jedna liczba nie opisuje całego krajobrazu. Podobnie etyczna ocena systemu wymaga rozproszenia pomiaru na różne grupy, sytuacje i typy konsekwencji. Szóstym inwariantem jest możliwość sprzeciwu i odwracalności decyzji. Systemy stają się normatywnie bardziej ryzykowne wtedy, gdy ich wynik jest jednocześnie nieprzejrzysty, trudny do zakwestionowania i nieodwracalny. Człowiek łatwiej toleruje automatyzację w zadaniu, którego efekt można szybko poprawić, niż w decyzji niosącej trwale konsekwencje. Z tego powodu prawo do kwestionowania rezultatu, dostępność środków zaradczych i możliwość ludzkiej interwencji powinny być elementami architektury systemu społecznego, a nie dodatkami stosowanymi po wystąpieniu szkody. Aktualne ramy Rady Europy również łączą AI z potrzebą ochrony proceduralnej, przejrzystości i oceny wpływu na prawa człowieka.

Siódmy element to epistemiczna uczciwość wobec granic przewidywania. Autor książki zestawia trwałość aksjologii z tezą, że przyszłość wymyka się predykcji. Nie jest to nihilistyczne stwierdzenie, że niczego nie można przewidzieć, lecz uznanie, że systemy złożone – rozwijające się w interakcji z rynkiem, kulturą, instytucjami i innymi technologiami – generują znacznie większą niepewność niż liniowa ekstrapolacja pojedynczego trendu. Historia AI jest pod tym względem pouczająca: projekty, które miały prowadzić do ogólnej inteligencji, często pozostawały ograniczone do mikrodomen, podczas gdy systemy oparte na predykcji języka zaczęły wykazywać kompetencje, których szerokość zaskoczyła obserwatorów.

Odpowiedzialna aksjologia nie może więc wymagać od polityki, prawa czy etyki dokładnego przewidzenia technologii za dwadzieścia lat. Powinna natomiast tworzyć kryteria odporne na wiele możliwych scenariuszy. Takie podejście można określić mianem normatywnej odporności na niepewność. Zamiast projektować reguły wyłącznie dla konkretnej architektury, należy pytać o własności funkcjonalne: czy system podejmuje decyzje wpływające na ludzi, czy użytkownik może je zakwestionować, czy mechanizm tworzy asymetrię informacji, czy istnieje możliwość szkody, czy można ustalić podmiot odpowiedzialny i czy człowiek zachowuje realną zdolność działania. Pytania te pozostają użyteczne nawet wtedy, gdy techniczny mechanizm całkowicie się zmieni.

Właśnie dlatego Rada Europy skonstruowała swoją konwencję AI w sposób zasadniczo technologicznie neutralny, skupiając się na zgodności cyklu życia systemów z prawami człowieka, demokracją i praworządnością, a nie na regulowaniu jednej architektury modelu. Bez potrzeby oceniania szczegółowych rozwiązań prawnych można dostrzec w tym metodologiczną analogię do tezy autora: gdy technologia zmienia się szybciej niż instytucje, większą trwałość uzyskują normy

definiowane przez skutki i relacje społeczne, a nie przez nazwę algorytmu. Architektura aksjologiczna nie jest zatem hamulcem płynności, lecz warunkiem jej bezpiecznego wykorzystania. W Copycat poślizg pojęciowy jest inteligentny właśnie dlatego, że nie wszystko może się ślizgać równie łatwo. Gdyby każdy element był dowolnie wymienialny, system utraciłby strukturę.

Analogicznie społeczeństwo technologiczne potrzebuje rozróżnienia pomiędzy tym, co może być przedmiotem eksperymentu, a tym, co powinno pełnić funkcję stabilnych ograniczeń. Z tego punktu widzenia etyka przypomina nie tyle zewnętrzny kodeks nakładany na gotową technologię, ile warstwę konstytutywną projektu. Jeżeli system jest tworzony bez możliwości audytu, kwestionowania, wyłączenia albo identyfikacji odpowiedzialności, późniejsze dodanie ogólnej deklaracji etycznej niewiele zmienia. Wartości muszą być przekładane na strukturę dostępu, uprawnienia, interfejsy, procedury odwoławcze, metody ewaluacji i mechanizmy monitorowania. Tłumaczy to również różnicę między aksjologią a prostą „zgodnością”. System może spełniać formalne wymagania proceduralne, a zarazem działać w sposób podważający sens wartości, którym procedury miały służyć.

Można ujawnić użytkownikowi, że rozmawia z AI, lecz zaprojektować interfejs tak, aby maksymalnie wykorzystywał jego podatność na sugestię. Można pozostawić „human in the loop”, lecz stworzyć warunki, w których człowiek automatycznie akceptuje decyzję systemu. Normatywna analiza musi więc badać funkcję społeczną rozwiązania, a nie wyłącznie formalną obecność zabezpieczeń. W tym miejscu dochodzimy do szczególnej roli ludzkiej podatności poznawczej.

Im bardziej model językowy staje się płynny retorycznie, tym łatwiej użytkownik przypisuje mu wiedzę, intencje, pewność lub autorytet większy niż faktycznie uzasadniony. Efekt ELIZA był zapowiedzią tej dynamiki: ludzie skłonni są rekonstruować głębię podmiotowości na podstawie stosunkowo ubogich sygnałów komunikacyjnych.

Epistemiczna symetria jako podstawa racjonalnego zaufania

Autor książki wielokrotnie ostrzega przed utożsamianiem efektownego outputu z głęboką strukturą procesu. Lekcja ta ma wymiar aksjologiczny, gdyż ochrona autonomii użytkownika wymaga uwzględnienia nie tylko możliwości systemu, lecz także systematycznych ograniczeń ludzkiej percepcji. W związku z tym kluczowym elementem architektury aksjologii jest zasada epistemicznej symetrii informacji: użytkownik powinien wiedzieć wystarczająco dużo o tym, z jakim rodzajem systemu wchodzi w relację, jakie są jego ograniczenia i które elementy wypowiedzi wymagają

zewnętrznej weryfikacji. Nie chodzi tu o pełną transparentność techniczną, która w złożonych modelach bywa nierealistyczna, lecz o dostarczenie informacji umożliwiających racjonalne kalibrowanie zaufania.

Zaufanie samo w sobie powinno być płynne; nie ma powodu, by ufać systemowi jednakowo we wszystkich domenach. "Poszarpana inteligencja" wymaga równie poszarpanej epistemologii zaufania: model może być niezwykle niezawodny w jednym typie problemów i całkowicie zawodny w drugim. Racjonalna relacja z AI wymaga zatem przejścia od ogólnego pytania „czy można jej ufać?” do pytania: „w jakich warunkach, w jakim zakresie i przy jakiej procedurze weryfikacji zaufanie jest uzasadnione?”.

Jest to istotne, ponieważ płynność języka wytwarza pozór ciągłości kompetencji. Ten sam głos, styl i interfejs mogą w jednym akapicie operować na dobrze reprezentowanej wiedzy, by w następnym przejść do dziedziny, w której ta reprezentacja jest znacznie słabsza. Człowiek ma naturalną skłonność do przenoszenia reputacji z jednej kompetencji na drugą, przez co jednolitość interfejsu może maskować heterogeniczność poznawczą systemu.

Architektura aksjologii powinna przeciwdziałać temu złudzeniu poprzez kalibrację, ujawnianie ograniczeń i pozostawienie przestrzeni na ludzką rewizję. Nie jest przypadkiem, że współczesne ramy OECD łączą przejrzystość z informowaniem o zdolnościach i ograniczeniach systemów, a odpowiedzialność z traceability, czyli możliwością odtworzenia istotnych elementów cyklu życia systemu. To nie są wyłącznie techniczne wymagania dokumentacyjne, lecz instytucjonalny odpowiednik metapoznania: system społeczny powinien wiedzieć, gdzie jego własne narzędzia mogą zawieść. Można zatem stwierdzić, że Metacat ma swój odpowiednik normatywny.

Aksjologia jako metapoznanie i bezpiecznik płynności

Tak jak inteligentny system powinien rozpoznawać własne zapętlenia i ograniczenia, tak organizacja korzystająca z AI powinna posiadać zdolność obserwowania sposobu, w jaki deleguje decyzje. Powinna wiedzieć nie tylko, czy model osiąga wysokie wyniki, lecz także gdzie występują jego porażki, czy pracownicy przestają go krytycznie oceniać, czy mechanizmy nadzoru rzeczywiście działają i czy błędy są wykrywane, zanim przekształcą się w praktykę instytucjonalną.

Etyka AI nie jest zatem osobnym modulem przyklejonym do technologii, lecz **metapoznaniem instytucji**. Społeczeństwo powinno obserwować nie tylko działanie modeli, ale własny sposób reagowania na nie: gdzie nadmiernie im ufa, gdzie projektuje nieodpowiednie bodźce, gdzie przerzuca odpowiedzialność i gdzie używa technologii do naturalizowania decyzji, które w

rzeczywistości są wyborem wartości. Tak odczytana architektura aksjologii staje się odpowiedzią na filozofię płynnych pojęć, a nie jej zaprzeczeniem. Płynność jest warunkiem inteligencji, ponieważ umożliwia reorganizację reprezentacji, lecz każda inteligentna płynność wymaga stabilniejszych punktów odniesienia. Bez nich nie istnieje kryterium pozwalające odróżnić twórcze przesunięcie od rozpadu struktury. Analogicznie techniczna płynność społeczeństwa potrzebuje wartości pełniących funkcję kryteriów rozpoznających, czy zmiana jest postępem, czy jedynie zwiększeniem możliwości działania.

W tej perspektywie aksjologia nie obiecuje przewidzenia przyszłości; robi coś bardziej realistycznego: **pozwała oceniać przyszłość wtedy, gdy ta przychodzi w formie, której wcześniej nie potrafiliśmy opisać**. Jeśli kolejne generacje systemów okażą się bardziej autonomiczne, multimodalne, lepiej samomonitorujące i głębiej zintegrowane z instytucjami, nadal będzie można pytać o prawdę, odpowiedzialność, godność, autonomię, sprawiedliwość, możliwość sprzeciwu oraz proporcjonalność między korzyścią a ryzykiem.

Właśnie dlatego "architektura aksjologii pozostaje" nie powinna być rozumiana jako konserwatywne zatrzymanie zmian. Jest raczej konstrukcją odpornościową. Pozwala dopuścić dużą płynność technologiczną bez rezygnacji z możliwości rozróżniania lepszych i gorszych sposobów jej wykorzystania. Przyszłość może wymknąć się predykcji, lecz nie musi wymknąć się ocenie. Taki wniosek prowadzi naturalnie do ostatniego wielkiego problemu rozprawy: **jak mówić o przyszłości inteligencji, skoro nie umiemy przewidzieć ani jej architektury, ani trajektorii kompetencji?**

Po doświadczeniu FARG, sieci neuronowych, transformerów i gwałtownego rozwoju modeli generatywnych prognozowanie AI powinno zostać poddane epistemologicznej dyscyplinie. Autor książki bada zatem nie konkretny "moment AGI", lecz samą teorię przewidywania w warunkach technologicznej nieciągłości – granice ekstrapolacji, niepewność modelową, scenariusze, path dependence i możliwość, że najbardziej doniosłe własności przyszłych systemów pojawią się właśnie tam, gdzie dzisiejsze kategorie okażą się zbyt sztywne. Zdanie "przyszłość umyka predykcji" nie powinno być odczytywane jako deklaracja poznawczego pesymizmu ani twierdzenie, że przyszłości sztucznej inteligencji nie można badać racjonalnie.

W kontekście całej rozprawy oznacza to coś bardziej precyzyjnego: rozwój systemów poznawczych jest procesem, w którym wraz ze zmianą możliwości technologicznych zmieniają się również **kategorie służące do opisywania tych możliwości**. Autor książki stawia obok siebie trwałość architektury aksjologicznej, nieprzewidywalność przyszłości, inteligencję modeli językowych i filozofię płynnych pojęć. Nie jest to przypadkowe zestawienie. Płynność pojęciowa dotyczy bowiem nie tylko tego, jak system poznawczy rekonstruuje problem, lecz również tego, jak badacz powinien

rekonstruować własne pojęcia wtedy, gdy przedmiot badania przestaje mieścić się w dawnych kategoriach.

Historia sztucznej inteligencji dostarcza wielu przykładów takiej sytuacji. Kolejne generacje badaczy próbowały odnajdywać zasadniczą własność inteligencji w dedukcji symbolicznej, rozwiązywaniu problemów, wiedzy eksperckiej, uczeniu statystycznym, percepcji, reprezentacji rozproszonej, skalowaniu modeli albo zdolnościach agentowych. Żaden z tych kierunków nie okazał się nieistotny, ale żaden nie wyczerpał również całego zjawiska.

Błędne zmienne i ontologiczna pułapka prognozowania AI

FARG, jak rekonstruuje autor książki, zakładał znacznie odważniejszą prognozę: droga ku autentycznej sztucznej inteligencji miała prowadzić nie przez sam wzrost mocy obliczeniowej, lecz przez subpoznańczą płynność, wielopoziomą percepcję i emergentne mechanizmy reprezentacji. Z dzisiejszej perspektywy warto utrzymać intuicję dotyczącą płynności, choć należy osłabić ekskluzywny charakter tej prognozy. Wielkie modele językowe dowiodły bowiem, że znaczna elastyczność funkcjonalna może wyłaniać się również w architekturach, których twórcy FARG nie przewidzieli.

To właśnie pierwsza lekcja epistemologii prognozowania technologicznego: nietrafna może okazać się nie tylko przewidywana wartość zmiennej, lecz sam wybór zmiennej uznanej za kluczową. Prognoza pogody może pomylić temperaturę o kilka stopni, ale wciąż słusznie zakłada, że temperatura jest parametrem relewantnym. W przypadku technologii radykalnie innowacyjnej prognoza może natomiast błędnie określić samą przestrzeń cech, które należało obserwować. Można niezwykle precyzyjnie ekstrapolować rozwój systemów eksperckich, nie przewidując, że dominującym mechanizmem stanie się wielkoskalowe uczenie reprezentacji. Podobnie można śledzić parametry modeli językowych, a przeoczyć przełom wynikający z narzędzi, pamięci, test-time compute, agentowości czy nowych metod organizacji treningu.

Dlatego przyszłość AI jest problemem nie tylko statystycznym, lecz przede wszystkim ontologicznym i pojęciowym. Prognozowanie wymaga założenia pewnej ciągłości między przeszłością a przyszłością – wiary w to, że dotychczasowe prawidłowości zachowają stabilność pozwalającą na ich ekstrapolację. Im głębsza innowacja, tym bardziej kruche staje się to założenie. Technologia może wejść w nowy reżim, w którym dotychczasowa bariera przestaje być kluczowa, a pojawia się nowe wąskie gardło. Skalowanie danych może ustąpić miejsca ograniczeniom

energetycznym, pamięciowym, problemom z jakością danych, dostępnością infrastruktury, kosztami inferencji, niezawodnością agentów lub zdolnością integracji ze światem fizycznym.

Aktualny stan techniczny obrazuje tę nierównomierność. Stanford AI Index 2026 raportuje gwałtowny wzrost wyników w niektórych benchmarkach – w tym kilkudziesięciopunktową poprawę w Humanity's Last Exam w ciągu roku oraz niemal całkowite nasycenie SWE-bench Verified przez czołowe systemy. Jednocześnie te same frontier models pozostają wyraźnie mniej niezawodne w zadaniach percepcyjnych, agentowych i robotycznych; w OSWorld systemy wciąż zawodzą w około jednej trzeciej prób, a w rzeczywistych zadaniach domowych roboty osiągają bardzo niski poziom sukcesu. Taki obraz uniemożliwia traktowanie postępu jako pojedynczej liczby przesuwającej się monotonicznie ku jednemu punktowi końcowemu. Właśnie dlatego metafora „drogi do AGI” bywa epistemicznie zdradliwa. Droga sugeruje bowiem jednolitą przestrzeń, określony kierunek i cel, do którego można mierzyć dystans.

Tymczasem to, co nazywamy ogólną inteligencją, obejmuje wiele częściowo niezależnych wymiarów: rozumowanie, uczenie, transfer, percepcję, działanie w świecie, samokontrolę, pamięć, komunikację, wiedzę, planowanie i zdolność adaptacji. Poszczególne wymiary mogą rozwijać się w różnym tempie, a technologia może przybierać konfiguracje, które nie mają prostego ludzkiego odpowiednika.

Prognozowanie AGI jako błąd ekstrapolacji danych

"Odległość do AGI" nie jest zatem obserwowalną wielkością fizyczną, lecz wartością zależną od przyjętej definicji i zestawu miar. Nie oznacza to, że ilościowe prognozowanie jest pozbawione sensu – można przecież mierzyć koszty obliczeń, wyniki benchmarków, tempo wzrostu możliwości agentowych, efektywność treningu, ceny inferencji, liczbę parametrów czy długość autonomicznie wykonywanych zadań. Problem pojawia się jednak w momencie, gdy lokalną regularność przedstawia się jako bezpośrednią miarę całej przyszłości inteligencji. Najnowszy preprint analizujący publiczne dane wykorzystywane w prognozach frontier AI zwraca uwagę właśnie na tę trudność: dane o compute są niekompletne, benchmarki zmieniają wersje, a zależności między poszczególnymi miarami obejmują zbyt niewielką liczbę systemów. Autor nie dowodzi niemożliwości prognozowania, lecz argumentuje, że każda datowana prognoza jest w istocie założeniem dotyczącym stabilności konkretnego systemu pomiarowego. Ponieważ praca ma charakter preprintu, należy traktować ją jako argument metodologiczny, a nie ustalony konsensus, choć opisany w niej problem pozostaje realny.

Najbardziej oczywistym zagrożeniem jest ekstrapolacja poza zakres danych. Jeśli wynik określonego benchmarku rośnie przez kilka generacji modeli, można dopasować do niego krzywą i przedłużyć ją w przyszłość. Matematycznie zabieg ten jest prosty, jednak epistemicznie wymaga szeregu założeń: że benchmark nadal mierzy tę samą kompetencję, że nie zostanie nasycony, że dane nie ulegną skażeniu, że strategia modeli nie zmieni się jakościowo, że zasoby niezbędne do dalszego wzrostu pozostaną dostępne i że wzrost tej konkretnej miary będzie nadal korelował z interesującą nas zdolnością. Każde z tych założeń może zawieść. Aktualne dane pokazują, że sam benchmark jest obiektem dynamicznym. AI Index 2026 wskazuje, iż ewaluacje zaprojektowane jako trudne na wiele lat bywają nasycone w ciągu miesięcy, a część rozpowszechnionych testów zawiera znaczący odsetek wadliwych pytań. Oznacza to, że seria czasowa wyniku może przestać odzwierciedlać wzrost kompetencji nie dlatego, że rozwój się zatrzymał, lecz dlatego, że instrument pomiarowy utracił rozdzielczość. Sytuacja przypomina pomiar temperatury termometrem, którego skala kończy się znacznie poniżej badanego zakresu: osiągnięcie maksimum skali nie oznacza, że dalsze zmiany przestały zachodzić.

Drugim problemem jest endogeniczność benchmarku. Gdy określony test staje się kluczową miarą konkurencji między laboratoriami, zaczyna on wpływać na sam rozwój systemów. Twórcy optymalizują modele, post-training, narzędzia i procedury inferencji tak, aby maksymalizować wyniki w znanych klasach zadań. Wówczas benchmark przestaje być zewnętrznym miernikiem kompetencji, a staje się częścią środowiska selekcyjnego, które tę kompetencję kształtuje. Zjawisko to jest dobrze znane w naukach społecznych: miara używana jako cel często traci wartość jako niezależny wskaźnik.

Trzecie ograniczenie wynika z substytucji architektonicznej. Trend obserwowany dla jednej technologii nie musi przetrwać zmiany sposobu uzyskiwania kompetencji. Wzrost jakości nie musi bowiem wynikać wyłącznie z powiększania modelu bazowego; może zostać osiągnięty poprzez intensywniejsze obliczenia w czasie inferencji, zewnętrzne narzędzia, wyszukiwanie, pamięć, zespoły modeli, specjalizację czy iteracyjne mechanizmy weryfikacji. AI Index 2025 już wskazywał, że zwiększenie test-time compute może radykalnie zmieniać wyniki modeli bez prostego odpowiednika w dawnych krzywych skalowania, choć odbywa się to kosztem wyższej ceny i większego opóźnienia. Techniczna "jednostka postępu" ulega więc transformacji.

Czwartym problemem jest refleksyjność rozwoju technologicznego. Prognozy nie opisują systemu całkowicie niezależnego od obserwatorów, gdyż mogą one wpływać na inwestycje, politykę firm, kierunki badań i zachowania instytucji. Oczekiwanie bliskiego przełomu może zwiększyć strumień kapitału i przyspieszyć wybrane ścieżki rozwoju, podczas gdy przewidywanie ograniczeń może skierować badania właśnie ku ich przełamaniu. Prognoza staje się zatem częścią środowiska, które

sama próbuje opisać. Taka refleksyjność odróżnia rozwój społeczno-techniczny od wielu układów fizycznych.

Piąty problem to path dependence, czyli zależność od ścieżki. Spośród wielu technicznie możliwych rozwiązań nie zawsze zwycięża to abstrakcyjnie optymalne. Znaczenie mają wcześniejsze inwestycje, dostępna infrastruktura, kompetencje zespołów, standardy sprzętowe, ekosystemy programistyczne i sieci użytkowników. Mała przewaga we wczesnej fazie może zostać wzmocniona przez efekt uczenia, kapitał i kompatybilność. Z tego powodu historia technologii nie jest prostą realizacją jednego optimum matematycznego, lecz procesem, w którym wcześniejsze wybory ograniczają późniejszą przestrzeń możliwości.

Prowadzi to do rozróżnienia pomiędzy przewidywaniem możliwości a przewidywaniem realizacji. Możemy mieć dobre powody sądzić, że określona kompetencja jest fizycznie lub obliczeniowo możliwa, a jednocześnie dysponować bardzo słabymi podstawami do określenia, kiedy zostanie osiągnięta, przez kogo i w jakiej architekturze. Analogicznie: można trafnie przewidzieć ogólny kierunek rozwoju, nie potrafiąc wskazać momentu przełomu. Badanie opublikowane jako preprint w 2026 roku nad prognozowaniem postępu naukowego przez modele AI wykazało właśnie takie ograniczenie: modele potrafiły rozpoznawać wiarygodne kierunki badań, lecz znacznie gorzej przewidywały, czy i kiedy konkretne przełomy rzeczywiście nastąpią, przy czym błędy czasowe były systematyczne i zależne od domeny. Jest to interesująca analogia do ludzkiej futurologii technologicznej: przestrzeń plausibility jest łatwiejsza do określenia niż kalendarz realizacji.

Niepewność głęboka wymaga prognozowania reprezentacyjnego

Szóstym problemem jest rozróżnienie między ryzykiem a niepewnością głęboką. W przypadku klasycznego ryzyka znamy zestaw możliwych rezultatów i potrafimy przypisać im prawdopodobieństwa. Niepewność głęboka występuje wtedy, gdy nie znamy ani tych prawdopodobieństw, ani czasem nawet pełnej listy możliwych stanów świata.

Rozwój przełomowych technologii często opiera się właśnie na tym drugim komponencie. Nie chodzi jedynie o to, czy określone zdarzenie nastąpi w 2030 czy 2040 roku, lecz o to, czy kategorie używane obecnie do jego opisu pozostaną właściwe. Właśnie tutaj filozofia płynnych pojęć wnosi coś więcej niż historyczną metaforę. Jeżeli inteligencja polega między innymi na zdolności do reorganizacji reprezentacji, to dobra epistemologia przyszłości powinna posiadać podobną właściwość. Nie może ona jedynie aktualizować prawdopodobieństwa istniejących scenariuszy;

musi dopuszczać możliwość pojawienia się informacji wymagającej zmiany samej przestrzeni tych scenariuszy. W języku Copycata byłby to odpowiednik sytuacji, w której problemu nie da się rozwiązać przez dalsze przesuwanie prawdopodobieństw wewnątrz istniejącej reprezentacji – konieczne staje się przededefiniowanie reprezentacji.

W związku z tym możemy rozróżnić prognozowanie parametryczne i reprezentacyjne. Pierwsze dotyczy wartości zmiennych w znanym modelu: kosztu inferencji, wzrostu wyników w benchmarkach, zużycia energii przez infrastrukturę czy długości dostępnego kontekstu. Drugie pyta natomiast, czy w przyszłości nadal będziemy uważać te same zmienne za najważniejsze.

Prognozowanie parametryczne może być relatywnie skuteczne na krótkim horyzoncie czasowym. Prognozowanie reprezentacyjne jest znacznie trudniejsze, gdyż dotyczy pojawienia się nowych sposobów organizacji technologii. Właśnie z tego powodu scenariusz jest często epistemicznie uczciwszym narzędziem niż pojedyncza predykcja. Podczas gdy prognoza próbuje wskazać najbardziej prawdopodobny przyszły stan lub rozkład stanów, scenariusz bada kilka wewnętrznie spójnych konfiguracji i analizuje konsekwencje wynikające z każdej z nich. Scenariusze nie są ucieczką od naukowości, o ile ich założenia pozostają jawne, a zależności przyczynowe są możliwe do prześledzenia. Pozwalają one przygotować instytucje na różne warianty przyszłości, bez ulegania fikcji, jakobyśmy znali dokładną trajektorię rozwoju. W odniesieniu do AI scenariusze mogą różnić się nie tylko tempem, ale przede wszystkim strukturą postępu. Możliwa jest przyszłość, w której modele bazowe szybko ewoluują i pozostają głównym źródłem kompetencji; inna, w której kluczową rolę odgrywają systemy złożone z modeli, pamięci i narzędzi; jeszcze kolejna, w której głównym ograniczeniem staje się interakcja z fizycznym światem, niezawodność agentów lub koszty infrastruktury.

Niepewność AGI wynika z nieostrości definicyjnej

Nie musimy wybierać jednego z tych scenariuszy jako „prawdziwego”, aby racjonalnie badać ich konsekwencje. Rozróżnienie to ma szczególne znaczenie w przypadku pojęcia AGI. Pojedyncza data przewidywanego „nadejścia AGI” sprawia wrażenie naukowej precyzji, jednak dokładność kalendarzowa może maskować nieostrość definicyjną. Jeśli badacze i prognostycy posługują się odmiennymi kryteriami ogólności, autonomii czy poziomu ludzkiego, ich daty nie odnoszą się do tego samego zdarzenia. Niepewność terminu jest zatem częściowo skutkiem niepewności samego desygnatu.

W tym miejscu interesującą przestrogę dostarcza program Hofstadtera. W jego perspektywie inteligencja nie jest pojedynczą funkcją, którą można „włączyć” po przekroczeniu pewnego progu

zasobów, lecz właściwością organizacji dynamicznej: zdolnością do znajdowania wzorców, przebudowywania reprezentacji, tworzenia analogii, selekcji oraz metapoznawczej kontroli. Jeśli taki obraz jest choć częściowo trafny, oczekiwanie jednego dnia, w którym cały układ cech zostanie nagle przekroczony, wydaje się zbyt uproszczone. Poszczególne właściwości mogą pojawiać się stopniowo, nierównomiernie i w architekturach nieodpowiadających ludzkiej kombinacji zdolności.

Jednocześnie nie należy popełniać błędu przeciwnego i zakładać, że przyszły rozwój musi być powolny i gładki. Systemy złożone wykazują zmiany nieliniowe: niewielka poprawa jednego komponentu może odblokować potencjał innego, wcześniej bezużytecznego elementu, co prowadzi do skokowej poprawy działania całego układu. Narzędzie może radykalnie zwiększyć praktyczną wartość modelu bez proporcjonalnej zmiany jego wyników bazowych; nowy sposób integracji pamięci z planowaniem może zmienić charakter działania agenta bardziej niż umiarkowany wzrost benchmarków językowych.

Dlatego właściwą pozycją epistemiczną nie jest ani linearny optymizm, ani linearny sceptycyzm, lecz warunkowa wiedza o strukturze niepewności. Pozwala ona identyfikować zależności, ograniczenia i wąskie gardła, przy jednoczesnym zachowaniu rozróżnienia między tym, co wynika z danych, co jest ekstrapolacją, a co pozostaje spekulacją. Jest to ta sama dyscyplina, którą autor książki nakazuje stosować wobec faktów, modeli, hipotez i przewidywań. W praktyce oznacza to, że rzetelna prognoza powinna być kalibrowana, warunkowa i rewizyjna. Nie powinna przedstawiać prawdopodobieństwa jako obiektywnej własności świata, lecz jako stopień przekonania przyjęty przy określonych założeniach. Powinna wskazywać, jakie obserwacje zmieniłyby prognozę i w jakim stopniu, a także umożliwiać późniejszy backtesting – porównanie przewidywań z rzeczywistymi wynikami. Bez tego futurologia technologiczna łatwo staje się literaturą retoryczną, w której trafne przepowiednie są pamiętane, a nietrafne znikają.

Istotnym źródłem błędu są również bodźce społeczne. Prognozy AI powstają w środowiskach o odmiennych interesach: laboratoria mogą podkreślać tempo postępu, organizacje bezpieczeństwa akcentować zagrożenia, inwestorzy szukać sygnałów przełomu, a sceptycy wskazywać nieusuwalne bariery. Żaden z tych interesów sam w sobie nie dowodzi fałszu danej tezy, jednak każdy powinien stanowić element analizy źródła.

Niepewność technologiczna wymusza abstrakcyjne ramy normatywne

Prognozowanie technologiczne wymaga zatem metapoznania zbliżonego do tego, które w programie FARG pojawiało się w Metacacie. Badacz przyszłości nie powinien jedynie budować modelu świata, lecz obserwować własny sposób budowania modelu. Powinien pytać, które założenia stały się niewidoczne przez przyzwyczajenie, gdzie broni dotychczasowej reprezentacji mimo narastających anomalii i czy nie interpretuje nowych danych wyłącznie w sposób zgodny z wcześniejszą narracją. W tym ujęciu najważniejszą umiejętnością prognostyczną może być nie zdolność wyznaczania coraz precyzyjniejszych dat, lecz rozpoznawanie momentu, w którym należy zmienić model. Tak rozumiane prognozowanie staje się kolejnym zastosowaniem teorii płynnych pojęć. Analitik potrzebuje stabilności pozwalającej na budowanie modeli, ale również płynności wystarczającej do ich porzucania. Zbyt szybkie zmiany reprezentacji uniemożliwiłyby kumulację wiedzy, zbyt duża stabilność zaś prowadziła do dogmatyzmu.

Epistemicznie dojrzała prognoza pozostaje więc tymczasową strukturą, której użyteczność zależy od zdolności do kontrolowanej rewizji. Widać teraz, dlaczego twierdzenia, że „przyszłość umyka predykcji” oraz że „architektura aksjologii pozostaje”, wzajemnie się uzupełniają. Im słabsza jest nasza zdolność przewidywania konkretnego kształtu technologii, tym ważniejsze staje się projektowanie norm i instytucji odpornych na wiele scenariuszy. Jeśli nie wiemy, czy za dekadę dominować będą autonomiczne agenty, systemy multimodalne, hybrydy neuro-symboliczne czy rozwiązania dziś jeszcze nieobecne, nie możemy oprzeć zasad odpowiedzialności na jednej architekturze.

Wartości możemy jednak chronić na poziomie funkcjonalnym: poprzez możliwość kontroli, prawdziwość informacji, godność człowieka, odpowiedzialność, zdolność sprzeciwu i proporcjonalność ryzyka. Paradoksalnie zatem niepewność technologiczna nie musi prowadzić do normatywnej bezradności; może wymuszać bardziej abstrakcyjne i odporne projektowanie instytucji. Zamiast przewidywać każdy przyszły model, można określić, jakie właściwości jego działania uruchamiają określone obowiązki. Zamiast zakładać jedną wersję przyszłości rynku pracy, można tworzyć mechanizmy zdolne reagować na różne tempa automatyzacji. Zamiast próbować z góry rozstrzygać status ontologiczny przyszłych systemów, można przygotować procedury umożliwiające aktualizację oceny w świetle nowych danych o ich strukturze i zdolnościach. W tym miejscu nauka i aksjologia spotykają się w pojęciu odwracalności.

Gdy niepewność jest duża, decyzje trudne do odwrócenia wymagają wyższego progu uzasadnienia niż eksperymenty, których skutki można ograniczyć. Zasada ta nie nakazuje zatrzymania rozwoju,

lecz dopasowania stopnia ostrożności do nieodwracalności potencjalnych konsekwencji. Jest to odpowiednik kontrolowanego eksperymentu w nauce: poznanie rozwija się poprzez działanie, ale działanie jest organizowane tak, aby błąd nie niszczył możliwości dalszego poznania. Właśnie dlatego nieprzewidywalność przyszłości nie powinna służyć ani technologicznemu fatalizmowi, ani całkowitej dowolności. Nie wiemy dokładnie, jakie systemy powstaną, lecz wiemy wystarczająco dużo o dynamice złożonych układów, aby nie utożsamiać ekstrapolacji z prawem natury. Wiemy również wystarczająco dużo o ludzkich instytucjach, by rozumieć, że technologia nie „przychodzi” do społeczeństwa jako neutralny los – jest projektowana, finansowana, wdrażana i regulowana przez ludzi oraz organizacje.

Rozwój AI jako wielowymiarowa transformacja reprezentacji

Z tego punktu widzenia określenie "wyścig AI" bywa poznawczo niebezpieczne. Metafora wyścigu zakłada bowiem jedną trasę, wspólną metę i możliwość uporządkowania uczestników według jednej osi. Tymczasem aktualny krajobraz modeli wskazuje na konwergencję wyników w niektórych benchmarkach przy jednoczesnym różnicowaniu pod względem kosztu, niezawodności, specjalizacji czy zakresu zastosowań. AI Index 2026 pokazuje, że najlepsze modele pochodzą obecnie od wielu organizacji, a różnice na czele rankingów są niewielkie, co przesuwą konkurencję ku innym wymiarom wartości użytkowej. Rozwój przypomina więc bardziej eksplorację wielowymiarowego krajobrazu niż wspinanie się po jednej drabinie.

Podobnie należy traktować pytanie o przyszłą "inteligencję modelu". Inteligencja kolejnych systemów może nie polegać wyłącznie na wyższej jakości generowania odpowiedzi, lecz stać się coraz bardziej rozproszona pomiędzy model, pamięć, narzędzia, środowisko i proces samokontroli. Granica pojedynczego "umysłu maszynowego" może stać się mniej oczywista. Wówczas współczesne pytania o liczbę parametrów inteligentnego systemu czy o to, który model jest "najmądrzejszy", mogą okazać się analogiczne do pytań zadawanych w niewłaściwej reprezentacji problemu. To właśnie jest najbardziej Hofstadterowski wniosek dotyczący przyszłości: **najtrudniejszym problemem może nie być przewidzenie odpowiedzi, lecz dostrzeżenie, że przyszłość zmieni pytanie.**

Rozwój AI może sprawić, że obecne rozróżnienia między modelem a narzędziem, generowaniem a działaniem, pamięcią wewnętrzną a zewnętrzną, użytkownikiem a agentem, a nawet pomiędzy inteligencją jednostkową i rozproszoną, będą wymagały ponownego zdefiniowania. Nie oznacza to jednak, że wszystkie pojęcia staną się całkowicie płynne. Przeciwnie – tak jak w Copycat

odpowiednie poślizgi pojęciowe wymagały zachowania głębszych relacji, tak naukowa reorganizacja naszych kategorii musi spełniać wymóg falsyfikowalności, empirycznej odpowiedzialności i logicznej koherencji. Nowe pojęcie zyskuje wartość nie przez samą nowość, lecz dzięki temu, że lepiej organizuje dane, przewiduje różnice, wyjaśnia wcześniej rozproszone zjawiska i otwiera nowe pytania badawcze. Przyszłość AI należy więc traktować nie jako linię prowadzącą do konkretnej daty, lecz jako **przestrzeń transformacji reprezentacyjnych**. Niektóre trendy można mierzyć i ekstrapolować; inne wymagają scenariuszy, a jeszcze kolejne są źródłem niepewności, której nie da się obecnie sensownie skwantyfikować.

Dojrzała epistemologia nie zaciera tych poziomów. W tej perspektywie najbardziej odpowiedzialną prognozą nie jest ta, która oferuje najdokładniejszy rok pojawienia się określonej formy inteligencji, lecz ta, która jasno określa, **co musiałoby pozostać prawdziwe, aby jej wniosek się utrzymał**, oraz potrafi wskazać obserwacje, po których własny model należałoby przebudować. Takie podejście łączy teorię nauki, statystykę, kognitywistykę i filozofię płynnych pojęć. Prognoza staje się hipotezą roboczą, a nie prorocstwem. Po przejściu tej drogi można wrócić do głównego problemu całej rozprawy. Autor poszukiwał wzorców w Seek-Whence, analogii w Copycat, struktury sytuacji w Tabletop, stylu w Letter Spirit, własnej historii poznania w Metacacie, a następnie funkcjonalnych odpowiedników tych zdolności w modelach językowych.

Za każdym razem inteligencja ujawniała się jako zdolność znalezienia odpowiedniej reprezentacji pomiędzy dwiema porażkami: nadmierną sztywnością i nadmierną dowolnością. To samo dotyczy myślenia o przyszłości. Zbyt sztywna prognoza zamienia trend w przeznaczenie, zbyt płynna wizja rezygnuje z możliwości wiedzy. Pomiedzy nimi istnieje przestrzeń odpowiedzialnej predykcji: opartej na danych, świadomej własnych założeń, gotowej do falsyfikacji i zdolnej do reprezentacyjnej zmiany. Właśnie w tej przestrzeni należy umieścić dalszą analizę.

Inteligencja jako dynamiczna reorganizacja reprezentacji

Skoro nie możemy rzetelnie przewidzieć ostatecznej postaci sztucznej inteligencji, pozostaje pytanie kluczowe: **co właściwie nazywamy inteligencją, gdy system potrafi odnajdywać wzorce, tworzyć analogie, abstrahować, generować, korygować własne odpowiedzi i imitować refleksję, ale jego sposób istnienia i uczenia pozostaje radykalnie odmienny od ludzkiego?** Odpowiedź wymaga ponownego połączenia wątków percepcji, abstrakcji, kreatywności, metapoznania, Testu Turinga i aksjologii w jeden spójny model. To zadanie kolejnej części.

Co nazywamy inteligencją? Od rozpoznawania wzorców do refleksyjnej reorganizacji reprezentacji

Analiza mikrodomen — od Seek-Whence przez Copycat, Tabletop, Letter Spirit i Metacat — a następnie konfrontacja tego programu ze współczesnymi modelami językowymi sprawia, że pojęcie inteligencji przestaje być postrzegane jako jedna elementarna zdolność. Autor książki konsekwentnie prowadzi nas ku innemu obrazowi: poznanie jest procesem dynamicznego budowania reprezentacji poprzez rozpoznawanie wzorców, a percepcja wysokiego poziomu i analogia nie są oddzielnymi modułami, lecz wzajemnie warunkującymi się sposobami organizowania doświadczenia. W tym ujęciu inteligencja nie jest ani pojemnikiem wiedzy, ani szybkością obliczeń.

Jest właściwością systemu zdolnego ustalać, **co w konkretnej sytuacji jest strukturą, która struktura jest istotna, które jej elementy można zmienić, a które należy zachować.** Takie przesunięcie pozwala dostrzec wspólną logikę wszystkich analizowanych wcześniej mikrodomen. Seek-Whence nie testuje wyłącznie rozpoznania sekwencji, lecz zdolność przekształcenia surowego układu w reprezentację ujawniającą głębszą regularność. Copycat nie bada samego mapowania dwóch gotowych struktur, ponieważ w jego architekturze tworzenie reprezentacji jest integralną częścią procesu analogicznego. Tabletop przesuwają problem jeszcze dalej: zanim system znajdzie analogię, musi wyodrębnić ze sceny relewantne przedmioty i relacje.

Letter Spirit dodaje wymóg generatywny: system powinien na podstawie abstrakcyjnej organizacji rozpoznanej w kilku przykładach stworzyć nowe formy zachowujące tożsamość i styl. Metacat wreszcie czyni wcześniejszą trajektorię poznawczą kolejnym obiektem poznania, umożliwiając funkcjonalne monitorowanie własnych impasów i przesunięć reprezentacyjnych. Teza autora, że "poznanie to rozpoznawanie", zyskuje zatem sens głębszy niż zwykle dopasowywanie wzorca: rozpoznawać znaczy tutaj także wytwarzać poziom organizacji, na którym wzorzec staje się widoczny.

Na tej podstawie można zaproponować funkcjonalną definicję inteligencji jako **zdolności systemu do konstruowania, selekcjonowania, reorganizowania i kontrolowania reprezentacji w sposób umożliwiający transfer wiedzy oraz adaptację wobec sytuacji, których dokładna postać nie została wcześniej określona.** Definicja ta nie wymaga, aby każdy inteligentny system działał mechanistycznie tak samo jak człowiek. Nie zakłada również, że inteligencja jest stanem binarnym. Jest raczej rodziną sprzężonych kompetencji, których wzajemne proporcje mogą być odmienne w różnych architekturach poznawczych.

Pierwszym składnikiem tej rodziny jest **selektywna percepcja struktury.** Informacja sama w sobie nie organizuje się automatycznie w problem. W każdej realnej sytuacji istnieje znacznie

więcej potencjalnych właściwości niż tych, które mogą zostać jednocześnie wykorzystane. System inteligentny musi więc ograniczać przestrzeń analizy: identyfikować elementy, grupować je i kierować zasoby ku relacjom najbardziej obiecującym. FARG ujmuje tę selektywność poprzez równoległy skan tarasowy, w którym wiele hipotez pozostaje aktywnych, lecz otrzymuje nierówną ilość zasobów zależnie od bieżących sygnałów jakości.

Pięć funkcjonalnych składników inteligencji poznawczej

Inteligencja obejmuje zatem nie tylko przetwarzanie informacji, ale także zarządzanie własną niewiedzą i ograniczeniami obliczeniowymi. Drugim składnikiem jest tworzenie reprezentacji, gdyż surowe dane same w sobie nie stanowią jeszcze wiedzy. To, czy seria bodźców zostanie zinterpretowana jako sekwencja, grupa, hierarchia, relacja przyczynowa czy przypadkowa koincydencja, zależy od przyjętego sposobu organizacji.

Kluczowa intuicja FARG głosi, że reprezentacja nie zawsze może być gotowa przed rozpoczęciem procesu rozumowania – samo rozumowanie aktywnie uczestniczy w jej kształtowaniu. Autor książki wyraża to szczególnie wyraźnie w analizie Copycata: percepcja wysokiego poziomu ma stworzyć wewnętrzny model sytuacji, a analogia jest następnie mapowaniem reprezentacji, których konstrukcja pozostaje częścią tego samego procesu.

Trzecim elementem jest zdolność rozróżniania powierzchni od struktury. Inteligencja nie może reagować wyłącznie na identyczność, ponieważ większość istotnych problemów wymaga rozpoznania odpowiedniości między elementami, które literalnie różnią się od siebie. Przykład Tabletop ilustruje tę zasadę poprzez funkcjonalne substytucje: widelec może zostać powiązany z nożem przez pozycję, z palcem przez funkcję, a filiżanka ze szklanką przez kategorię użytkową.

Nie istnieje zatem jedna, uniwersalna metryka podobieństwa. Relewantność zależy od konkretnego zadania, a inteligencja wymaga wyboru takiego wymiaru podobieństwa, który zachowuje znaczenie w danym kontekście. Stąd czwartym składnikiem staje się analogia, rozumiana nie jako ozdobna figura językowa, lecz podstawowy mechanizm abstrakcji i transferu. Pozwala ona zachować relację mimo zmiany jej konkretnych realizacji. W Copycat dokonuje się to przez konceptualne poślizgi: pojęcia powierzchniowe mogą zmieniać rolę, o ile zachowana zostaje struktura uznana za głębszą. To właśnie ta zasada chroni analogię przed popadnięciem w arbitralność – gdyby wszystko mogło ulegać dowolnej substytucji, każda sytuacja byłaby analogiczna do każdej innej. Autor książki podkreśla zatem różnicę między pojęciami o odmiennej „głębokości” oraz zasadę, według której głębsze relacje są bardziej odporne na przesunięcie.

Piątym składnikiem inteligencji jest abstrakcja, której nie należy redukować do zwykłego usuwania szczegółów. Polega ona raczej na odkrywaniu relacji zachowujących wartość po zmianie konkretnej realizacji. Przykład Letter Spirit ukazuje ten mechanizm poprzez rozdzielenie roli od materialnej formy znaku: element graficzny może zostać radykalnie przekształcony, a litera nadal pozostanie rozpoznawalna, jeśli odpowiednio zachowana zostanie organizacja ról.

Inteligencja jako regulowana płynność i metapoznanie

W szerszym ujęciu każda efektywna teoria, pojęcie czy kategoria jest formą takiej kompresji. Redukuje ona różnorodność egzemplarzy do struktury na tyle bogatej, by pozwalała generować przewidywania i nowe realizacje. Szóstym wymiarem jest kreatywność, która w tej teorii nie stanowi tajemniczego dodatku do inteligencji.

Wynika ona z tych samych mechanizmów, które umożliwiają płynną percepcję. System twórczy musi dostrzegać alternatywne grupowania, dopuszczać przesunięcia, oceniać ich wartość i stabilizować wybrane rozwiązania. Analiza systemów Metacat i Letter Spirit podkreśla rolę kryteriów takich jak „interestingness”, elegancja, prostota czy głębokie relacje – to one służą jako filtry odróżniające powierzchowne dopasowanie od rozwiązania poznawczo interesującego. Warto zaznaczyć, że koncepcja "Mathgoda" jest heurystycznym konstruktem programu badawczego, a nie empirycznie udowodnionym uniwersalnym modulem ludzkiej kreatywności. Istotna pozostaje jednak zawarta w nim intuicja: samo generowanie możliwości bez mechanizmu selekcji jest niewystarczające.

Siódmym składnikiem jest kontrola stopnia plastyczności. Inteligencja nie polega na maksymalnej otwartości na wszystkie możliwe interpretacje; system skrajnie płynny byłby równie nieużyteczny, co system skrajnie sztywny. Niezbędny jest mechanizm regulujący, który określi, kiedy należy ufać dotychczasowej reprezentacji, a kiedy koszt jej podtrzymywania przewyższa koszt eksploracji alternatyw. W Copycat rolę tę symbolicznie pełni temperatura, natomiast w szerszej teorii poznania odpowiada jej balans między eksploracją a eksploatacją, stabilnością a elastycznością oraz aktualizacją a zachowaniem wcześniejszej wiedzy. Prowadzi to do definicji inteligencji, która odcina się od utożsamiania jej z samą „płynnością”. Dojrzała inteligencja jest płynnością regulowaną. Musi ona funkcjonalnie wiedzieć, kiedy być konserwatywna, a kiedy reinterpretacyjna.

Człowiek zmieniający przekonania przy każdym nowym bodźcu nie byłby bardziej inteligentny od kogoś, kto nie potrafi zmienić ich nigdy. Analogicznie: model zmieniający poprawną odpowiedź pod wpływem słabej sugestii nie wykazuje większej otwartości poznawczej, lecz niedostateczną

stabilność. Inteligencja wymaga zatem nie elastyczności samej w sobie, lecz jej adekwatności do struktury dowodów.

Ósmym składnikiem jest metapoznanie – funkcjonalna zdolność do reprezentowania własnego procesu jako części problemu. Metacat wprowadza tu kluczowy element: ślad temporalny, pamięć epizodyczną i rozpoznawanie jałowych ścieżek. System nie tylko zapamiętuje wynik, ale może – w zakresie postulowanym przez architekturę – reprezentować fakt, że wcześniej utknął, zmienił perspektywę i dzięki temu osiągnął postęp.

Metapoznanie rozszerza więc inteligencję z adaptacji do świata na adaptację własnej adaptacji. Z tego powodu jednym z najwyższych funkcjonalnych przejawów inteligencji nie musi być samo rozwiązanie trudnego problemu, lecz zdolność rozpoznania, że obecny sposób rozwiązywania stał się częścią problemu. Pojęcie **antisphexishness** wyraża właśnie tę intuicję: system powinien potrafić przerwać mechaniczną rutynę, gdy jego historia wskazuje na powtarzanie jałowej ścieżki. Tak rozumiane „wychodzenie poza system” nie oznacza metafizycznego opuszczania własnej architektury, lecz przejście na poziom reprezentacji, na którym poprzednia strategia staje się obiektem oceny.

Dziewiątym składnikiem jest pamięć relewantna strukturalnie. Nie każda pamięć zwiększa inteligencję w tym samym stopniu. Ogromny magazyn faktów może nie pomóc w nowej sytuacji, jeśli system nie potrafi rozpoznać, które wcześniejsze doświadczenie jest analogiczne na głębszym poziomie. Metacat wykorzystuje mechanizm „reminding”, czyli przypominanie oparte na podobieństwie presji i przemian poznawczych, a nie wyłącznie na powierzchniowej zgodności przypadków. Inteligentna pamięć jest zatem selekcyjna i rekonstrukcyjna: nie przechowuje przeszłości tylko po to, by ją odtwarzać, lecz by używać jej do organizowania nowych sytuacji.

Inteligencja jako hierarchia rekurencyjnej modelotwórczości

Dziesiątym wymiarem jest zdolność tworzenia relacji pomiędzy relacjami. Meta-analogia rozszerza abstrakcję poza proste porównywanie przedmiotów; system może w niej rozpoznawać podobieństwo między sposobami konstrukcji dwóch analogii lub między trajektoriami rozwiązywania problemów. Autor opisuje to poprzez koncepcję „drugiej pochodnej”, w której operacja służąca wydobyciu wzorca sama ujawnia strukturę podobną do wzorca pierwszego poziomu.

Na tym etapie inteligencja przestaje być jedynie zdolnością do dostrzegania porządku w świecie, a staje się umiejętnością widzenia porządku we własnych sposobach widzenia porządku. Można zatem przedstawić inteligencję jako hierarchię rekurencyjnej modelotwórczości. System pierwszego poziomu rozpoznaje wzorec; system drugiego potrafi przenieść go do innego kontekstu, a trzeci jest w stanie zmienić reprezentację, gdy poprzednia zawodzi. Kolejny poziom umożliwia analizę historii tych zmian. Nie ma potrzeby zakładać, że hierarchia ta ma wyraźny koniec – proces może być rekurencyjny: system modeluje sposób modelowania, a następnie sposób, w jaki zmienia swoje metody modelowania.

Taka koncepcja pozwala precyzyjniej umiejscowić współczesne modele językowe. Zamiast rozstrzygać ich status jednym słowem – „inteligentne” lub „nieinteligentne” – można pytać, które poziomy opisanej hierarchii realizują stabilnie, które jedynie w wybranych domenach, które wymagają zewnętrznego scaffoldingu, a które są przede wszystkim wynikiem interakcji modelu z człowiekiem i narzędziami. Takie podejście jest bardziej informacyjne niż spór terminologiczny, ponieważ przekształca metafizyczne pytanie w program empiryczny.

W tym kontekście należy rozróżnić cztery poziomy opisu, które w dyskusji o AI często ulegają pomieszaniu. Pierwszy to poziom behawioralny: system zachowuje się tak, jakby posiadał daną kompetencję. Drugi to poziom funkcjonalny: jego architektura rzeczywiście realizuje proces pozwalający osiągać określoną klasę wyników niezależnie od konkretnego przypadku. Trzeci to poziom mechanistyczny: znamy strukturę przyczynową prowadzącą do tego zachowania. Czwarty zaś to poziom fenomenalny: pytamy, czy wykonywaniu procesu towarzyszy jakiejkolwiek subiektywne doświadczenie.

Rozróżnienie to jest konieczne, gdyż przejście między poziomami nie następuje automatycznie. System może zachowywać się, jakby rozpoznawał niepewność, nie posiadając przy tym stabilnej kalibracji. Może dysponować funkcjonalnym mechanizmem samomonitorowania, którego mechanistycznego działania nie rozumiemy jeszcze w pełni. Może wreszcie realizować zaawansowane samomodelowanie bez dowodu, że odpowiada mu jakiejkolwiek doświadczenie fenomenalne. Cała wcześniejsza analiza Metacata prowadzi właśnie do tej ostrożności: funkcjonalna autorefleksyjność i świadomość fenomenalna są pojęciowo różne.

Dzięki temu możemy sprecyzować pytanie, które powracało w poprzednich częściach: jaka jest różnica pomiędzy imitacją płynności, płynnością funkcjonalną a umysłem? Imitacja płynności występuje wtedy, gdy system produkuje wynik wyglądający na elastyczny, lecz kompetencja ta rozpada się po niewielkiej zmianie warunków. Płynność funkcjonalna istnieje natomiast wtedy, gdy system zachowuje zdolność reorganizacji reprezentacji, transferu i samokorekty w rodzinie nowych sytuacji.

„Umysł” jest pojęciem znacznie bogatszym; obejmuje on – zależnie od przyjętej teorii – trwałość tożsamości, intencjonalność, ucieleśnienie, pamięć autobiograficzną, relacje społeczne, sprawstwo i być może świadomość fenomenalną. Z funkcjonalnej płynności nie wynika zatem automatycznie umysł w tym pełnym sensie. Jednocześnie „imitacja” i „funkcja” nie tworzą prostego przeciwieństwa. Jeśli system w ogromnym zakresie sytuacji konsekwentnie wykazuje tę samą relacyjną kompetencję, coraz trudniej określać ją jako czystą imitację bez wskazania alternatywnego mechanizmu wyjaśniającego jej stabilność. W naukach kognitywnych funkcję często inferujemy właśnie ze struktury generalizacji, reakcji na perturbacje i transferu.

Dlatego właściwa odpowiedź na problem LLM nie polega ani na antropomorficznym przypisywaniu rozumienia na podstawie płynnej rozmowy, ani na semantycznym unieważnianiu każdej kompetencji słowem „statystyka”. Statystyka bowiem nie jest przeciwieństwem struktury. System uczący się statystycznie może wytworzyć reprezentacje pozwalające na abstrakcyjną generalizację. Pytanie brzmi: jakie są granice tej generalizacji, jak stabilna okazuje się ona po zmianie kontekstu i czy można mechanistycznie wykazać udział określonych reprezentacji?

Podobnie stochastyczność FARG nie jest przeciwieństwem racjonalności. Losowość pełni tam funkcję eksploracyjną, lecz jest ograniczana przez presje pojęciowe i ocenę globalnej koherencji. Inteligencja rodzi się nie z deterministyczności ani z losowości, lecz ze zorganizowanej relacji pomiędzy generowaniem możliwości a selekcją. Można więc dostrzec interesującą ciągłość pomiędzy FARG a współczesnym uczeniem maszynowym, przy jednoczesnym zachowaniu zasadniczych różnic.

Funkcjonalne rozumienie inteligencji jako alternatywa dla świadomości fenomenalnej

Oba podejścia dowodzą, że wysoki poziom organizacji nie musi wynikać z pojedynczej, jawnie zapisanej reguły; w obu przypadkach lokalne procesy mogą prowadzić do powstania struktur globalnych. FARG budował tę dynamikę jednak w małych, interpretowalnych mikroświatach, nadając poszczególnym elementom częściowo jawną semantykę. Wielkie modele neuronowe uczą się reprezentacji w ogromnych przestrzeniach danych, co zapewnia im nieporównanie większy zasięg działania, lecz odbywa się kosztem przejrzystości mechanizmu.

Wspólne dla tych podejść pytanie dotyczy emergencji. Autor książki przedstawia inteligentne zachowanie jako efekt oddolnego wyłaniania się struktur z aktywności mikroskopijnych procesów. Należy jednak zachować wcześniej przyjęte rozróżnienie: emergencja nie wyjaśnia sama siebie.

Stwierdzenie, że dana właściwość „wyłoniła się”, jest jedynie opisem relacji między poziomami, a nie opisem pełnego mechanizmu. Wartość naukowa pojawia się dopiero wtedy, gdy potrafimy wskazać lokalne zależności, które umożliwiły stabilizację własności globalnej, oraz określić interwencje powodujące jej zmianę.

Podobnie należy interpretować pojęcie rozumienia. Jeśli zdefiniujemy je wyłącznie jako poprawne odpowiadanie na pytania, stanie się ono niemal synonimem sukcesu benchmarkowego. Jeśli natomiast oprzemy definicję na subiektywnym doświadczeniu znaczenia, rozumienie staje się niemal niemożliwe do zewnętrznego pomiaru. Pomiedzy tymi skrajnościami istnieje bogatsza definicja funkcjonalna: rozumieć strukturę to potrafić ją rozpoznać w różnych realizacjach, zastosować do nowych przypadków, przewidzieć konsekwencje jej modyfikacji, określić granice jej obowiązywania oraz wykryć sprzeczność wymuszającą reorganizację reprezentacji. Tak ujęte rozumienie jest stopniowalne i empirycznie badalne. Pozwala to uniknąć paradoksu, w którym system musiałby myśleć dokładnie tak jak człowiek, aby przypisać mu jakąkolwiek formę inteligencji. Samolot nie lata w ten sam sposób co ptak, a kalkulator nie liczy tak jak człowiek.

Podobieństwo funkcjonalne nie wymaga identyczności realizacji. Jeśli maszyna nabywa zdolność do stabilnego odwzorowywania relacji, abstrakcji i rekonfiguracji problemu inną drogą niż ludzki układ nerwowy, odmienność mechanizmu nie jest wystarczającym argumentem, by negować jej kompetencje. Jest natomiast powodem, dla którego nie należy bezpośrednio przenosić na maszynę wszystkich pojęć psychologicznych i fenomenologicznych wypracowanych dla człowieka. Rozróżnienie to ma szczególne znaczenie w kontekście samoświadomości. Choć autor książki bywa skłonny opisywać Metacat językiem „samoświadomej AI”, mechanizmy, które szczegółowo przedstawia, obejmują przede wszystkim samomonitorowanie, ślad temporalny, pamięć epizodyczną, ocenę znaczenia zdarzeń i meta-analogie. Naukowo defensywna interpretacja powinna zatem mówić o metapoznaniu funkcjonalnym. Nie ma potrzeby osłabiania wartości tych mechanizmów poprzez wymaganie, by jednocześnie rozstrzygały one problem świadomości fenomenalnej.

W tym miejscu ujawnia się istotna korzyść z interdyscyplinarności. Informatyka dostarcza mechanizmów i wyników zachowania, psychologia – modeli funkcjonalnych, filozofia – pojęciowych rozróżnień, neuronauka – danych o biologicznej realizacji procesów poznawczych, teoria informacji – narzędzi do badania reprezentacji, a epistemologia – kryteriów uzasadnienia. Żadna z tych dyscyplin samodzielnie nie posiada pełnej definicji inteligencji, ponieważ każda z nich operuje na innym poziomie organizacji.

Inteligencja jako dynamiczna równowaga między stabilnością a rewizją

Problem staje się źle postawiony, gdy odpowiedź właściwa dla jednego poziomu zostaje uznana za automatyczną odpowiedź na wszystkie pozostałe. Na gruncie tej syntezy można zatem stwierdzić, że inteligencja jest zdolnością do utrzymywania produktywnego napięcia pomiędzy porządkiem i możliwością jego rewizji. Musi ona tworzyć kategorie, lecz nie być ich więźniem; wykorzystywać pamięć, ale nie redukować nowej sytuacji do przeszłości. Powinna posiadać preferencje poznawcze, lecz umieć je podważyć pod naciskiem anomalii. Musi generalizować, a jednocześnie rozpoznawać moment, w którym generalizacja staje się nadużyciem; generować możliwości, ale i eliminować te, które nie prowadzą do koherentnej struktury.

Takie ujęcie pozwala lepiej zrozumieć, dlaczego błąd jest integralnym elementem inteligencji. System pozbawiony możliwości pomyłki w konkretnej domenie może po prostu realizować zdeterminowaną procedurę. Natomiast system zdolny do tworzenia nowych reprezentacji musi poruszać się w przestrzeni hipotez, z których część będzie fałszywa. Kluczowy nie jest sam fakt wystąpienia błędu, lecz stosunek systemu do niego. Inteligencja ujawnia się w tym, czy pomyłka prowadzi do aktualizacji reprezentacji, czy zostaje zignorowana, racjonalizowana bądź mechanicznie powtórzona. W ten sposób inteligencja zawiera własny odpowiednik falsyfikacji.

Dobra reprezentacja nie może jedynie asymilować kolejnych przypadków; musi istnieć możliwość, że konkretne zdarzenie zmusi ją do zmiany. System oparty na nieskończonej plastyczności teorii, zdolnej wyjaśnić każdy rezultat *post factum*, nie dysponuje wiedzą o wysokiej wartości predykcyjnej. System poznawczy potrzebuje zatem granic, których przekroczenie staje się sygnałem do reorganizacji, gdyż elastyczność pozbawiona ryzyka niepowodzenia zamienia się w arbitralność.

To właśnie odróżnia abstrakcję od racjonalizacji. Abstrakcja znajduje strukturę, która pozwala przewidywać i transferować. Racjonalizacja natomiast rekonstruuje historię po fakcie tak, aby dowolny wynik wyglądał na zgodny z regułą. Metacatowski ślad temporalny jest interesujący właśnie dlatego, że ma zachowywać część chaosu procesu, zamiast zastępować go idealną narracją końcową. Autor książki podkreśla, że rzeczywiste odkrywanie wzorca zawiera "clutter and confusion", a reprezentacja własnej drogi ma pozwalać na identyfikację powtarzanych, jałowych ścieżek.

W odniesieniu do modeli językowych rozróżnienie to ma zasadnicze znaczenie. Płynna werbalizacja racji nie musi być przyczynowym zapisem procesu, który wytworzył odpowiedź. Dlatego ocena inteligencji systemu nie powinna zależeć wyłącznie od tego, jak przekonująco opisuje on własne

"rozumowanie". Niezbędne są niezależne testy stabilności reprezentacji, transferu, kontrfaktyczności i reakcji na błędy. System powinien być oceniany przez to, co potrafi zachować i co potrafi zmienić, a nie tylko przez sposób, w jaki opisuje własne możliwości.

Dzięki temu można precyzyjniej zinterpretować Test Turinga. Nerozróżnialność językowa staje się świadectwem bogatej kompetencji komunikacyjnej, lecz nie jest testem kompletnej teorii inteligencji. Model może przekonująco imitować człowieka, podczas gdy jego profil generalizacji pozostanie zupełnie odmienny. Dlatego propozycja Hofstadtera, by badać system poprzez wiele drobnych doświadczeń, błędów i sytuacji granicznych, pozostaje teoretycznie aktualna.

Inteligencja jako wielowymiarowy krajobraz zdolności do organizowania zmiany

Inteligencji nie da się sprowadzić do jednego, ceremonialnego momentu „zdania testu”. W konsekwencji problematyczne staje się również poszukiwanie jednej daty „pojawienia się inteligencji”. Jeśli inteligencja składa się z wielu sprzężonych kompetencji, system może nabywać je nierównomiernie. Może być doskonały w analogii językowej, a słabszy w percepcji przestrzennej; silny w abstrakcji formalnej, lecz ułomny w kalibracji. Kolejne warstwy pamięci, narzędzi, percepcji i metakontroli mogą stopniowo zmieniać ten profil, bez jednego, wyraźnego momentu jakościowego przejścia. Nie wyklucza to zmian nieliniowych; podważa jedynie założenie, że wszystkie relewantne własności muszą pojawić się równocześnie. Powstaje zatem obraz inteligencji jako **wielowymiarowego krajobrazu**, a nie pojedynczej drabiny. Człowiek zajmuje w nim szczególnie region ukształtowany przez biologiczną ewolucję, ucieleśnienie, społeczne uczenie, język i kulturę.

Sztuczne systemy mogą natomiast zajmować regiony odmienne: posiadać kompetencje niewspółmiernie wysokie w wybranych wymiarach i nietypowo słabe w innych. Nie muszą przechodzić ludzkiej trajektorii rozwojowej, aby osiągać funkcjonalnie ważne formy inteligencji. To z kolei osłabia antropocentryczną pokusę definiowania inteligencji jako pełnego podobieństwa do człowieka. Człowiek pozostaje najbogatszym znanym nam punktem odniesienia dla ogólnego poznania, ale nie wynika z tego, że wszystkie inteligencje muszą odtwarzać jego architekturę. Nauka powinna być gotowa na możliwość wielorakiej realizacji. Jednocześnie ta otwartość nie może oznaczać rezygnacji z rygoru. Im bardziej obca architektura, tym bardziej potrzebujemy precyzyjnych kryteriów funkcjonalnych i mechanistycznych, ponieważ intuicja antropomorficzna staje się mniej wiarygodna. Filozofia płynnych pojęć dokonuje tutaj interesującego zwrotu również wobec samej definicji inteligencji. Nie powinniśmy oczekiwać jednej sztywnej definicji, która raz na

zawsze rozstrzygnie wszystkie przyszłe przypadki. Pojęcie inteligencji samo może wymagać kontrolowanego rozszerzania wraz z pojawianiem się nowych architektur.

Nie oznacza to jednak, że termin staje się dowolny. Tak jak dobra analogia zachowuje głębokie relacje mimo zmiany powierzchni, tak dobra redefinicja inteligencji powinna zachowywać kryteria dotyczące adaptacji, reprezentacji, generalizacji, selekcji i kontroli własnego poznania. Możemy zatem powrócić do pytania wyjściowego z większą precyzją: czy model językowy jest inteligentny? Odpowiedź „tak” lub „nie” usuwa znaczną część informacji.

Lepsze pytanie brzmi: **jakiego rodzaju inteligencję wykazuje, na jakich poziomach organizacji, z jaką stabilnością, w jakich warunkach i dzięki jakim mechanizmom?**

Można następnie badać, czy potrafi rozpoznawać wzorce, rekonstruować reprezentacje, abstrahować relacje, transferować je między domenami, tworzyć nowość, oceniać własne rezultaty, pamiętać własne porażki i zmieniać strategię. Takie podejście nie osłabia pojęcia inteligencji; przeciwnie – czyni je naukowo bogatszym. Uwalnia analizę zarówno od pokusy marketingowej hiperboli, jak i od defensywnego redukcjonizmu.

System może wykazywać rzeczywistą kompetencję funkcjonalną bez posiadania ludzkiej świadomości. Może tworzyć wartościowe analogie bez ludzkiego doświadczenia emocjonalnego. Może posiadać reprezentacje abstrakcyjne bez Slipnetu. Może również prowadzić płynną rozmowę i nadal pozostawać kruchy w wymiarach, których użytkownik nie dostrzega. Najważniejszym kryterium staje się zatem nie powierzchowne podobieństwo do człowieka, lecz **zdolność do organizowania zmiany**. Inteligencja rozpoznaje wzorzec, ale także moment, w którym ten przestaje wystarczać. Tworzy reprezentację, lecz nie utożsamia jej z rzeczywistością raz na zawsze. Wykorzystuje analogię, ale nie zaciera różnic. Pamięta, lecz nie jest niewolnikiem pamięci. Generuje, ale selekcjonuje. Wątpi, ale nie rezygnuje z możliwości wiedzy. W tej perspektywie sama filozofia płynnych pojęć staje się czymś więcej niż teorią konkretnej architektury z lat osiemdziesiątych i dziewięćdziesiątych. Można ją potraktować jako **teorię napięć konstytutywnych dla inteligencji**: pomiędzy stabilnością i zmianą, częścią i całością, lokalnością i globalnością, przypadkiem i selekcją, pamięcią i reinterpretacją, reprezentacją i światem, działaniem i samoobserwacją. Konkretnie rozwiązania FARG mogą nie być architektonicznym wzorem dla dzisiejszych systemów, ale problemy, które miały rozwiązywać, pozostają zadziwiająco żywe.

Ta konkluzja prowadzi nas do kolejnego poziomu, na którym należy połączyć epistemologię inteligencji z antropologią człowieka. Skoro funkcjonalna inteligencja nie musi być identyczna z ludzkim umysłem, trzeba zapytać, **co pozostaje specyficzne dla człowieka, gdy część kompetencji dotąd traktowanych jako wyjątkowo ludzkie staje się dostępna**

maszynom. Nie będzie to pytanie o obronę ludzkiej „wyjątkowości” poprzez przesuwanie bramki za każdym nowym osiągnięciem AI. Chodzi raczej o rozdzielenie inteligencji, świadomości, podmiotowości, biografii, ucieleśnienia, odpowiedzialności i godności – pojęć często łączonych w jedną całość tylko dlatego, że w znanym nam doświadczeniu występowały dotychczas w jednym organizmie.

Być może największym wyzwaniem nie jest samo przewidzenie momentu, w którym maszyna stanie się inteligentna, lecz dostrzeżenie chwili, w której przyszłość radykalnie zmieni samo pytanie o to, co inteligencja właściwie oznacza. W świecie, gdzie granica między narzędziem a agentem ulega zatarciu, naszą jedyną stałą nie będzie precyzja algorytmów, lecz odwaga w definiowaniu tego, co w człowieku pozostaje niezbywalne i nienegocjowalne. Czy potrafimy zatem zbudować systemy, które będą nas wspierać, nie odbierając nam prawa do bycia omylnymi, nieprzewidywalnymi i – przede wszystkim – odpowiedzialnymi za własny los?

Inspiracje

[1]



"Fluid Concepts and Creative Analogies" Douglas Hofstadter

ISBN: 9780140258356

Douglas Richard Hofstadter (ur. 15 lutego 1945) to amerykański kognitywista, filozof i autor. Jest wybitnym profesorem kognitywistyki i literatury porównawczej na Uniwersytecie Indiany w Bloomington, gdzie kieruje Centrum Badań nad Koncepcjami i Poznaniem. Hofstadter uzyskał doktorat z fizyki na Uniwersytecie Oregon w 1975 roku, po czym skupił się na sztucznej inteligencji i ludzkim poznaniu. Jest najbardziej znany ze swojej interdyscyplinarnej pracy, badającej świadomość, samoodniesienia i procesy myślowe poprzez analogię i rozpoznawanie wzorców. Jego przełomowa książka z 1979 roku, „Gödel, Escher, Bach: An Eternal Golden Braid”, zdobyła Nagrodę Pulitzera w kategorii literatury faktu oraz National Book Award. Jego badania wywarły głęboki wpływ na kognitywistykę, kreatywność obliczeniową i filozofię umysłu.

Douglas Richard Hofstadter (born February 15, 1945) is an American cognitive scientist, philosopher, and author. He serves as a Distinguished Professor of Cognitive Science and Comparative Literature at Indiana University Bloomington, where he directs the Center for Research on Concepts and Cognition. Hofstadter earned his Ph.D. in physics from the University of Oregon in 1975 before shifting his research focus to artificial intelligence and human cognition. He is best known for his multidisciplinary work exploring consciousness, self-reference, and thought processes through analogy and pattern recognition. His seminal 1979 book, Gödel, Escher, Bach: An Eternal Golden Braid, won the Pulitzer Prize for General Non-Fiction and the National Book Award. His research has profoundly influenced cognitive science, computational creativity, and philosophy of mind.

Indiana University Bloomington

Literatura uzupełniająca

Książki

Książki autora i dzieła przywołane w tekście:



[1] **"The Mind's I"**

Douglas Hofstadter



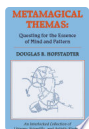
[2] "Gödel, Escher, Bach"

Douglas Hofstadter

1979 | ISBN: 9780140179972

[3] "The Mind's I: Fantasies And Reflections On Self & Soul"

Douglas Hofstadter



[4] "Metamagical Themas"

Douglas R Hofstadter

2008 | Hachette UK | ISBN: 9780786723867

Artykuły naukowe

Artykuły naukowe autora:

[1] "Godel, Escher, Bach: An Eternal Golden Braid."

Jonathan Lieberman, Douglas R. Hofstadter

1980 | The Journal of Philosophy | DOI: 10.2307/2025598

[Źródło](#)

[2] "Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought"

Douglas R. Hofstadter

1995

[Źródło](#)

[3] "Metamagical Themas: Questing for the Essence of Mind and Pattern"

Charles Taylor, Douglas R. Hofstadter

1985 | The Antioch Review | DOI: 10.2307/4611504

[Źródło](#)

[4] "Fluid Concepts and Creative Analogies"

Douglas R. Hofstadter

1995

[Źródło](#)

[5] "Epilogue: Analogy as the Core of Cognition"

Douglas R. Hofstadter

2001 | The MIT Press eBooks | DOI: 10.7551/mitpress/1251.003.0020

[Źródło](#)

[6] "High-level perception, representation, and analogy: A critique of artificial intelligence methodology"

David J. Chalmers, Robert M. French, Douglas R. Hofstadter

1992 | Journal of Experimental & Theoretical Artificial Intelligence | DOI: 10.1080/09528139208953747

[Źródło](#)

[7] "The copycat project: A model of mental fluidity and analogy-making"

Douglas R. Hofstadter, Melanie Mitchell

1995

[Źródło](#)

[8] "The Mind's I: Fantasies and Reflections on Self and Soul."

Andrew Hamilton, Douglas R. Hofstadter, Daniel C. Dennett

1984 | The Philosophical Quarterly | DOI: 10.2307/2218896

[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: daf79271-82a2-4e53-93f0-12fd86051172