

Poza Testem Turinga: profil poznawczy i problem przypisywania zasług w świetle koncepcji Douglasa Hofstadtera



AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje utratę mocy diagnostycznej Testu Turinga w obliczu rozwoju wielkich modeli językowych (LLM), które potrafią skutecznie imitować ludzkie zachowania konwersacyjne, nie posiadając przy tym ludzkiej architektury poznawczej. Autor wskazuje na wyniki badań, w których modele takie jak GPT-4.5 osiągają wysoki poziom nierozróżnialności od człowieka, co przesuwając ciężar debaty z pytania o samą inteligencję na analizę mechanizmów jej wytwarzania. W odpowiedzi na ten problem przywołano koncepcje Douglasa Hofstadtera i program FARG, proponując zastąpienie pojedynczego testu tzw. eksperymentami rozpraszającymi (scattering experiments). Podejście to ma na celu badanie struktury reakcji systemu w zróżnicowanych, nietypowych sytuacjach, co pozwala lepiej zrozumieć rzeczywisty profil poznawczy AI i odróżnić powierzchowną imitację od głębokiej kompetencji.

The article analyzes the loss of diagnostic power of the Turing Test in the face of the development of Large Language Models (LLMs), which can effectively imitate human conversational behaviors without possessing a human cognitive architecture. The author points to research results in which models such as GPT-4.5 achieve a high level of indistinguishability from humans, shifting the weight of the debate from the question of intelligence itself to an analysis of the mechanisms of its production. In response to this problem, Douglas Hofstadter's concepts and the FARG program are cited, proposing the replacement of a single test with so-called scattering experiments. This approach aims to study the structure of the system's reactions in diverse, atypical situations, allowing for a better understanding of AI's actual cognitive profile and distinguishing superficial imitation from deep competence.

Współczesne modele językowe (LLM) doprowadziły do kryzysu diagnostycznego Testu Turinga: fakt, że maszyna potrafi skutecznie imitować człowieka w krótkiej konwersacji, przestał być dowodem inteligencji, a stał się jedynie potwierdzeniem sprawności behawioralnej. Autor tekstu postuluje radykalne przejście od oceny efektu końcowego do wielopoziomowej analizy mechanistycznej i tzw. „eksperymentów rozpraszających” (scattering experiments) w duchu Douglasa Hofstadtera. Zamiast pytać, czy system „zaliczył” test, należy badać jego profil poznawczy poprzez testy kontrfaktyczne, badanie stabilności kompetencji oraz analizę procesów rekonfiguracji reprezentacji. Kluczowym wyzwaniem staje się wówczas problem 'credit assignment' – precyzyjne rozdzielanie przyczynowej autonomii procesu obliczeniowego od odpowiedzialności normatywnej i moralnej. Tekst dowodzi, że prawdziwa inteligencja nie objawia się w bezbłędnym wyniku, lecz w specyficznej dynamice radzenia sobie z porażką, zdolności do samokorekty oraz sposobie organizowania przestrzeni możliwości, co pozwala oddzielić techniczną sprawczość od podmiotowości.

SŁOWA KLUCZOWE

Test Turinga

profil poznawczy

problem przypisywania zasług

Douglas Hofstadter

LLM

eksperymenty rozpraszające

scattering experiments

interpretowalność mechanistyczna

testy kontrfaktyczne

kompozycjonalność

epistemiczna autonomia

out-of-distribution generalization

credit assignment

architektura aksjologiczna

emergencja poznawcza

Utrata mocy diagnostycznej Testu Turinga przez LLM

Test Turinga po epoce modeli językowych: od imitacji człowieka do eksperymentów rozpraszających

Test Turinga należy do tych idei w historii nauki, których znaczenie ewoluuje wraz z rozwojem technologii. Gdy Alan Turing w 1950 roku zaproponował „grę w naśladownictwo”, nie dążył do sformułowania laboratoryjnej definicji świadomości ani udowodnienia ontologicznej równoważności człowieka i maszyny. Przeciwnie, starał się ominąć jałowy spór o to, „czy maszyny mogą myśleć”, zastępując go pytaniem operacyjnym dotyczącym zachowania. Według Turinga, jeśli

kompetentny sędzia w trakcie interakcji tekstowej nie potrafi niezawodnie odróżnić maszyny od człowieka, wówczas przynajmniej część debaty o maszynowym myśleniu musi zostać przeformułowana. Turing świadomie przesunął ciężar z metafizycznej definicji „myślenia” na empirycznie dostępną organizację zachowania.

Historyczna doniosłość tego posunięcia polegała na metodologicznym minimalizmie. Turing nie twierdził, że nierozróżnialność konwersacyjna ujawnia kompletną strukturę umysłu; zaproponował raczej kryterium, które pozwoliło przeciąć nieskończoną dyskusję o wewnętrznej istocie maszyny poprzez odwołanie do obserwowalnych kompetencji. Co szczególnie interesujące z dzisiejszej perspektywy, jego koncepcja nie ograniczała się do czystego imitowania przeciętnego człowieka.

Turing przewidywał, że maszyna może celowo popełniać błędy, uczyć się i generować zachowania, których programista nie jest w stanie szczegółowo przewidzieć. Już u źródeł testu obecne było zatem napięcie pomiędzy imitacją a inteligencją: sprawny uczestnik gry nie musi być człowiekiem ani działać identycznie jak on, lecz musi umieć wytworzyć taki profil zachowania, który wystarczająco długo podtrzymuje interpretację „to rozmówca ludzki”. Przez dziesięciolecia test zachowywał znaczenie głównie dlatego, że kompetencja językowa stanowiła dla maszyn barierę niemal nie do przejścia. Swobodna rozmowa wymaga bowiem nie tylko składni, lecz także wiedzy o świecie, elastyczności semantycznej, poczucia humoru, umiejętności zmiany tematu, rozumienia niedopowiedzeń, korekty własnych wypowiedzi i dostosowania się do stylu rozmówcy.

Dopóki systemy komputerowe zawodziły już na pierwszych etapach takiej interakcji, Test Turinga dysponował wysoką siłą diagnostyczną. Sytuacja zmieniła się zasadniczo wraz z rozwojem wielkich modeli językowych. To, co wcześniej było rzadkim osiągnięciem demonstracyjnym, stało się elementem codziennego interfejsu systemów używanych przez miliony ludzi. W maju 2026 roku Jones i Bergen opublikowali w **Proceedings of the National Academy of Sciences** wyniki kontrolowanych, prerejestrowanych badań trójstronnego Testu Turinga.

Uczestnicy prowadzili jednocześnie pięciominutową rozmowę z człowiekiem oraz z jednym z systemów, po czym wskazywali, który z nich jest ludzki. W badaniu GPT-4.5, odpowiednio instruowany do przyjęcia ludzkiej osoby, osiągał poziom identyfikacji jako człowiek przekraczający przypadek; w wersji badania raportowanej wcześniej przez autorów model z personą był wskazywany jako człowiek w 73% przypadków. Autorzy artykułu stwierdzili, że kilka współczesnych modeli osiągnęło co najmniej 50-procentowy wskaźnik „przejścia” standardowego testu. Wynik ten ma znaczenie historyczne, lecz jego interpretacja wymaga dyscypliny. Nie dowodzi on, że model posiada świadomość fenomenalną ani że wszystkie jego procesy poznawcze są funkcjonalnie równoważne ludzkim. Dowodzi czegoś bardziej ograniczonego, a zarazem kluczowego: w określonych warunkach krótka konwersacja tekstowa przestała wystarczać wielu

ludziom do niezawodnego odróżnienia niektórych modeli od człowieka. Innymi słowy, kryterium imitacyjnej nierozróżnialności, które przez lata uchodziło za wyjątkowo wysoką poprzeczkę, zostało w konkretnych warunkach eksperymentalnych osiągnięte.

Właśnie tutaj pojawia się problem istotny dla filozofii Hofstadtera. Jeśli test będący symboliczną granicą maszynowej inteligencji można przejść za pomocą systemu, którego wewnętrzna organizacja pozostaje bardzo odległa od ludzkiej architektury poznawczej, przestaje on dostarczać wystarczających informacji o mechanizmie. Nie staje się bezwartościowy, lecz zmienia swoją funkcję: zamiast pytać, czy system „jest inteligentny”, pozwala badać, w jakim stopniu potrafi on wytworzyć przekonujący profil ludzkiego zachowania językowego i jakie sygnały ludzie wykorzystują do rozpoznawania podmiotowości rozmówcy.

To przesunięcie było zapowiedziane w samym programie FARG. Hofstadter proponuje odejście od jednego ceremonialnego sprawdzianu na rzecz czegoś, co można określić jako *scattering experiments* – eksperymentów rozpraszających system po wielu drobnych, nietypowych i wzajemnie zróżnicowanych sytuacjach. Istotą tej koncepcji nie jest uzyskanie pojedynczego progu „zaliczenia”, lecz obserwowanie struktury reakcji: tego, gdzie system zachowuje płynność, gdzie ulega blokadzie, jak reaguje na kontrfaktyczne przekształcenia, czy potrafi zmieniać interpretację i czy popełniane przez niego błędy układają się w poznawczo sensowny profil. Takie podejście jest bardziej zgodne ze współczesnym obrazem „poszarpanej inteligencji”. AI Index 2026 pokazuje, że modele potrafią osiągać wyniki niezwykle wysokie w jednych domenach i pozostawać zaskakująco kruche w innych. Granica kompetencji nie przesuwają się więc jak równy front, lecz przybiera postać nieregularnego krajobrazu, w którym sprawność w konkursowej matematyce może współistnieć z dużo słabszymi wynikami w zadaniach percepcyjnych, które człowiek uważa za banalne. Jeśli tak wygląda profil systemu, pojedynczy test konwersacyjny nie może dostarczyć jego pełnego opisu poznawczego. Należy więc odróżnić trzy poziomy ewaluacji.

Przejście od efektu behawioralnego do analizy mechanizmów i stabilności

Pierwszy poziom to **produkt behawioralny**: ocena tego, czy odpowiedź wydaje się sensowna, ludzka, twórcza lub poprawna. Drugi to **stabilność kompetencji**: sprawdzenie, czy podobna jakość utrzymuje się po zmianie sformułowania, domeny, modalności czy warunków kontekstowych. Trzeci to **mechanizm przyczynowy**, czyli analiza tego, jakie reprezentacje i procesy rzeczywiście doprowadziły do wyniku.

Klasyczny Test Turinga operuje głównie na poziomie pierwszym, jednak współczesna kognitywistyka AI wymaga równoczesnego badania wszystkich trzech. W tym miejscu ujawnia się zasadnicza przewaga Hofstadterowskiego sposobu myślenia. Copycat nie miał jedynie odpowiedzieć `wyz`; jego celem było ujawnienie, jak zmieniają się konkurencyjne reprezentacje, które pojęcia stają się dostępne i jaki rodzaj impasu prowadzi do reorganizacji. Metacat z kolei miał przechowywać ślad własnego procesu, rozpoznawać zapętlenia i budować meta-analogie. Z tej perspektywy inteligentne zachowanie nie jest pojedynczym sukcesem, lecz **charakterystyczną dynamiką dochodzenia do sukcesów i porażek**. Test powinien zatem dążyć do odtworzenia tej dynamiki, a nie jedynie weryfikować efekt końcowy.

Współczesna interpretowalność mechanistyczna rozwija podobną intuicję, choć stosuje inne metody. Badania z 2025 roku wykazały możliwość budowania częściowych grafów atrybucji, które wskazują, jakie wewnętrzne cechy i zależności miały przyczynowy udział w wygenerowaniu konkretnego wyniku. Autorzy tych prac podkreślają jednak, że obecne narzędzia odkrywają jedynie fragment obliczeń i nie dają pełnego obrazu działania modelu. Jest to podejście odmienne od czystego behawioryzmu: zachowanie pozostaje punktem wyjścia, ale celem staje się rekonstrukcja drogi prowadzącej od reprezentacji do odpowiedzi. Można powiedzieć, że w klasycznym Teście Turinga "wnętrze skrzynki" zostaje celowo zawieszone, natomiast we współczesnej interpretowalności to właśnie wnętrze staje się przedmiotem badania.

Nie ma w tym sprzeczności, gdyż oba podejścia odpowiadają na inne pytania. Test behawioralny pyta o to, co system potrafi zrobić w interakcji; interpretowalność mechanistyczna pyta, jak to robi. Pełniejsza teoria inteligencji wymaga obu poziomów – zachowanie bez mechanizmu sprzyja antropomorfizacji, a mechanizm bez zachowania nie pozwala rozstrzygnąć, czy dana reprezentacja ma funkcjonalne znaczenie.

Kluczowe stają się zatem **testy kontrfaktyczne**. Zamiast pytać model wyłącznie o wiedzę zgodną z rzeczywistością, można zmienić jedną podstawową regułę i badać, czy system potrafi konsekwentnie przeliczyć wynikający z niej świat. Jeżeli na potrzeby lokalnego zadania `z` przestaje być ostatnią literą alfabetu lub relacja "większy od" zostaje zdefiniowana odwrotnie, system powinien nie tylko powtórzyć instrukcję, lecz utrzymywać konsekwencje nowej reguły przez cały proces rozumowania. Taki eksperyment jest bliższy Copycat niż klasycznemu quizowi wiedzy, ponieważ bada zdolność do rekonfiguracji reprezentacji. Test kontrfaktyczny ogranicza ponadto prostą przewagę pamięci: jeśli zadanie jest skonstruowane syntetycznie, model nie znajdzie dokładnego odpowiednika w danych treningowych.

System musi wówczas zbudować tymczasową reprezentację nowego mikroświata. Oczywiście może przy tym korzystać z wyuczonych heurystyk ogólnych, lecz właśnie to jest pożądane. Test ma badać

nie czystość od przeszłej wiedzy, lecz możliwość zastosowania znanych mechanizmów w konfiguracji, której wcześniej nie widziano. Innym wymiarem scattering experiment powinna być **zmiana powierzchni przy zachowaniu struktury**. Ten sam problem można zapisać za pomocą liter, liczb, obrazów, opowieści, schematu przestrzennego lub relacji społecznej.

Jeżeli system naprawdę uchwycił relację, powinien w pewnym stopniu zachować kompetencję mimo wymiany nośnika. Pełna inwariancja nie jest realistycznym wymaganiem, gdyż różne modalności niosą odmienne trudności; kluczowe jest jednak mierzenie gradientu spadku. Powolne pogarszanie się generalizacji ma inne znaczenie niż gwałtowne załamanie po kosmetycznej zmianie opisu. Podobną rolę powinny pełnić testy **parafraz i odwracania perspektywy**. W Tabletop widać, że relacja "po lewej" zależy od punktu widzenia; system może więc otrzymywać tę samą sytuację opisaną z perspektywy różnych uczestników. Jeśli jego reprezentacja jest relacyjna, powinien rekonstruować odpowiednią transformację. Jeśli zaś opiera się na powierzchniowych wzorcach językowych, niewielka zmiana osoby gramatycznej lub kolejności opisu może wywołać nieproporcjonalny błąd. Trzeci rodzaj prób powinien badać **pamięć struktury przez czas**.

Model może poprawnie zaakceptować lokalną regułę na początku rozmowy, by po kilkunastu krokach wrócić do dominujących regularności treningowych. W takim przypadku pierwszy wynik daje fałszywe wrażenie głębszego przyswojenia reguły. Dopiero dłuższa trajektoria pokazuje, czy nowa reprezentacja pozostaje aktywna wystarczająco długo, aby sterować wieloetapowym rozumowaniem.

Kompozycjonalność i epistemiczna autonomia jako mierniki inteligencji

Czwarty wymiar dotyczy kompozycjonalności. System nie powinien jedynie opanować dwóch reguł oddzielnie, lecz umieć połączyć je w nową kombinację. Człowiek rozumiejący operacje A i B często potrafi przewidzieć skutki sekwencji $B(A(x))$ bez wcześniejszego treningu na dokładnie takim układzie. Jeśli model działa wyłącznie w ramach znanych konfiguracji, jego kompetencja może być jedynie rozbudowanym rozpoznawaniem wzorców. Jeżeli natomiast sprawnie składa relacje w nowe struktury, argument za istnieniem bardziej abstrakcyjnej reprezentacji staje się silniejszy.

Piąty wymiar obejmuje kalibrację i rozpoznawanie niepewności. System może podać błędną odpowiedź, lecz poprawnie wskazać jej niską wiarygodność; może też odpowiedzieć trafnie, ale z bardzo niską pewnością. Z punktu widzenia praktycznego i epistemicznego przypadki te nie są równoważne. Model zdolny do wiarygodnego odróżniania sytuacji stabilnych od kruchych może

funkcjonować w szerszym systemie decyzyjnym znacznie bezpieczniej niż model o podobnej średniej dokładności, który nie potrafi przewidzieć własnych porażek. Metacat wskazywał właśnie na taką potrzebę: rezultat poznawczy powinien być oceniany nie tylko zewnątrz, ale i przez mechanizmy wewnętrznej kontroli. Autor zakładał możliwość przechowywania znaczących wydarzeń, wykrywania powtarzania nieskutecznego schematu oraz wykorzystywania własnej historii jako materiału dla przyszłego rozumowania. Współczesna ewaluacja powinna zatem badać, czy model odróżnia trafne rozumowanie od przypadkowego trafienia i czy informacja o niepewności realnie wpływa na jego dalszą strategię.

Szóstym wymiarem jest konfrontacja z własnym błędem. Proste pytanie „czy jesteś pewien?” jest niewystarczające, gdyż model może zmienić odpowiedź samym faktem zakwestionowania jej przez użytkownika. Bardziej informatywne jest dostarczenie sprzecznego dowodu o różnej jakości i obserwowanie, czy system różnicuje wiarygodność źródeł, czy jedynie konformistycznie dostosowuje się do ostatniej wiadomości. Właściwa samokorekta wymaga równoważenia dwóch ryzyk: uporczywego trwania przy błędzie oraz zbyt łatwego porzucania poprawnej odpowiedzi.

Prowadzi to do szerszego problemu epistemicznej autonomii. Inteligencja nie polega wyłącznie na szybkim reagowaniu na nowe dane; system musi oceniać jakość informacji i stopień, w jakim powinny one modyfikować jego aktualny model. Człowiek również bywa albo zbyt dogmatyczny, albo zbyt podatny na sugestię. Dobry test AI powinien więc badać nie tylko elastyczność, ale właściwą regulację elastyczności. W języku FARG byłoby to pytanie o to, czy temperatura rośnie wtedy, gdy rzeczywiście powinna, i czy stabilizacja następuje dopiero w momencie osiągnięcia dostatecznej globalnej koherencji.

Od behawioralnych testów do interwencji mechanistycznych

Kolejny rodzaj prób powinien obejmować out-of-distribution generalization, czyli generalizację poza typowy rozkład treningowy. To jedno z najbardziej nadużywanych pojęć w dyskusji o AI; trudno bowiem precyzyjnie określić, co dla systemu rzeczywiście znajduje się „poza rozkładem”, gdy zbiory danych uczących są ogromne. Z tego powodu eksperymenty powinny opierać się na proceduralnie generowanych środowiskach, sztucznych ontologiach i kombinacjach reguł tworzonych doraźnie na potrzeby testu. Ich wartość nie polega na absolutnym odcięciu systemu od wcześniejszej wiedzy, lecz na precyzyjnej kontroli dystansu między treningiem a zadaniem. Hofstadterowskie mikrodomeny odzyskują tutaj zaskakującą aktualność.

To, co w epoce systemów symbolicznych mogło jawić się jako ograniczenie programu badawczego, dziś może stać się zaletą metodologiczną. Sztucznie skonstruowany mikroświat daje badaczowi pełną kontrolę nad regułami i możliwość dokładnego generowania kontrprzykładów. Nie jest on modelem całej inteligencji, lecz doskonałym narzędziem do weryfikacji pojedynczych mechanizmów generalizacji. Kluczowe jest jednak połączenie tej kontroli ze skalą współczesnych modeli. Podczas gdy FARG badał płynność w małym świecie, co pozwalało mu śledzić proces, LLM operują w świecie ogromnym, lecz ich wewnętrzne procesy pozostają częściowo nieprzejryste. Współczesny program badawczy może spróbować połączyć oba podejścia: wykorzystać potężny model ogólny, ale umieścić go w sztucznych środowiskach diagnostycznych, gdzie wiadomo dokładnie, jaka abstrakcja jest wymagana i jakie rodzaje skrótów poznawczych zostały wyeliminowane.

Dopełnieniem tych działań powinny być interwencje mechanistyczne. Samo znalezienie korelacji między określoną aktywacją a zachowaniem nie dowodzi, że dana aktywacja przyczynowo odpowiada za kompetencję. Niezbędne są eksperymenty, w których reprezentację wzmacnia się, osłabia lub usuwa, by następnie obserwować zmianę odpowiedzi. Właśnie takie podejście odróżnia interpretowalność opisową od przyczynowej. Współczesne prace nad mechanistyczną interpretacją coraz częściej dążą do przejścia od identyfikowania „cech” ku rekonstrukcji obwodów i ścieżek wpływu. W perspektywie filozoficznej jest to próba odpowiedzi na klasyczny problem wielorakiej realizacji: dwa systemy mogą zachowywać się podobnie, mimo skrajnie różnych mechanizmów. Jeśli interesuje nas jedynie skuteczność narzędzia, wystarczy podobieństwo produktu. Jeśli jednak pytamy o naturę inteligencji, musimy wiedzieć, czy systemy osiągają efekt poprzez zbliżone reprezentacje, czy przez odmienne mechanizmy funkcjonalnie konwergentne. Turing świadomie zawiesił to pytanie, gdyż w 1950 roku było ono niemal niemożliwe do rozstrzygnięcia empirycznie.

Dzisiaj sytuacja wygląda inaczej. Choć nie rozumiemy jeszcze modeli dostatecznie dobrze, dysponujemy narzędziami pozwalającymi przynajmniej częściowo badać ich wewnętrzne reprezentacje. W efekcie epistemologiczny kompromis Testu Turinga – „oceniamy zachowanie, bo wewnątrz pozostaje niedostępne” – staje się mniej konieczny. Nie oznacza to jednak, że eksperymenty behawioralne tracą na znaczeniu. Wręcz przeciwnie: analiza mechanistyczna bez funkcjonalnego testowania może prowadzić do przeceniania roli lokalnych struktur. Badacz może zidentyfikować cechę powiązaną z konkretną treścią, lecz dopiero szerokie eksperymenty wykażą, czy odpowiada ona za stabilną generalizację. Konieczny jest zatem ruch w obu kierunkach: od zachowania do mechanizmu i od mechanizmu z powrotem do zachowania.

Od egzaminu katalogowego do mapowania profilu poznawczego

Właśnie to może stanowić współczesną wersję Hofstadterowskiego scattering experiment. Nie byłby to pojedynczy egzamin, lecz wielopoziomowa procedura badawcza obejmująca: konwersację, perturbację formy, kontrfaktyczność, transfer, test kompozycyjalny, kalibrację, analizę reakcji na błąd, generalizację poza rozkład, badanie wewnętrznych reprezentacji oraz eksperyment przyczynowy. Zamiast jednego wyniku otrzymalibyśmy profil poznawczy systemu. Takie podejście odpowiada również problemowi benchmark saturation. AI Index 2026 podkreśla, że modele coraz szybciej nasycają benchmarki projektowane jako bardzo trudne, a część popularnych zbiorów testowych zawiera liczne wadliwe pytania. Oznacza to, że stały bank pytań szybko traci wartość diagnostyczną, szczególnie gdy modele lub procesy treningowe mogą dostosować się do jego struktury. Test inteligencji musi więc coraz częściej przypominać eksperyment naukowy, a nie egzamin katalogowy. W tym miejscu pojawia się jeden z najbardziej paradoksalnych skutków rozwoju AI.

Im lepiej modele radzą sobie ze standardowymi testami, tym mniej informacyjne staje się samo "zaliczenie". Wysoki wynik jest ważnym osiągnięciem technicznym, ale po przekroczeniu pewnego poziomu badacz musi przejść do pytań bardziej precyzyjnych: jakie komponenty odpowiadają za wynik, czy kompetencja jest stabilna, czy istnieją klasy systematycznych porażek i jak system zachowuje się po modyfikacji struktury zadania. Tak należy również interpretować sukces współczesnych modeli w Teście Turinga. Zamiast traktować go jako koniec debaty o inteligencji, należałoby uznać go za moment wyczerpania jednego narzędzia diagnostycznego. Test nie został obalony; spełnił swoją historyczną funkcję. Pokazał, że pewien typ ludzkiej kompetencji komunikacyjnej jest możliwy do odwzorowania przez systemy sztuczne na poziomie wystarczającym do wprowadzenia człowieka w błąd.

Właśnie dlatego potrzebujemy teraz testów następnej generacji. Hofstadterowski punkt widzenia pozwala określić, czego powinny one szukać. Nie chodzi o "dziwną zagadkę", której maszyna jeszcze nie zna – taki test zostałby szybko włączony do kolejnego zbioru treningowego. Chodzi o badanie zdolności tworzenia i przebudowy reprezentacji. Zadanie powinno być zaprojektowane tak, aby jego rozwiązanie wymagało wykrycia nowego inwariantu, porzucenia wcześniejszej interpretacji, utrzymania kontrfaktycznej reguły albo rozpoznania analogii między odległymi domenami. Jeszcze silniejszy test powinien wymagać rozpoznania własnej porażki reprezentacyjnej. System otrzymuje zadanie, w którym pierwsza strategia prowadzi do impasu, jednak nie informujemy go, jaka reprezentacja jest właściwa. Obserwujemy wtedy, czy potrafi wykryć stagnację, zwiększyć

eksplorację, sformułować alternatywę i wyjaśnić, dlaczego poprzednia perspektywa była niewystarczająca. To byłby współczesny odpowiednik przejścia od Copycata do Metacata.

Nawet taki eksperyment nie rozstrzygałby jednak problemu świadomości. System może posiadać głęboką zdolność adaptacyjnego samomonitorowania bez fenomenalnego doświadczenia. Testy funkcjonalne powinny więc pozostać testami funkcjonalnymi. Jednym z największych błędów debaty o AI jest przeskakowanie od wysokiej kompetencji do twierdzeń ontologicznych. Z faktu, że system potrafi powiedzieć "zmieniłem strategię, ponieważ poprzednia prowadziła do sprzeczności", nie wynika automatycznie, że doświadczył frustracji, wglądu czy satysfakcji. Równie ważny jest błąd odwrotny: brak dowodu świadomości nie uprawnia do traktowania wszystkich kompetencji systemu jako pozornych. Można uznać rzeczywistą funkcjonalną zdolność do abstrakcji, planowania czy metakontroli, nie rozstrzygając zarazem pytania o fenomenalność. Nauka potrzebuje pojęć pośrednich, które pozwolą opisywać maszynę bez konieczności wybierania pomiędzy dwiema skrajnościami: "to tylko automat" i "to już osoba". Właśnie temu służy architektura pojęciowa budowana w dotychczasowych częściach.

Płynność, analogia, poślizg pojęciowy, zmiana reprezentacji, hierarchiczne grupowanie, funkcjonalne substytucje, samomonitorowanie i meta-analogia są pojęciami bogatszymi niż prosty binarny sąd "inteligentne/nieinteligentne". Pozwalają opisywać stopień i typ kompetencji bez przyjmowania przedwczesnej tezy ontologicznej. Test Turinga po epoce LLM powinien więc zostać odczytany nie jako nieważny relikw, ale jako pierwszy etap historii operacjonalizacji inteligencji. W 1950 roku istotnym pytaniem było, czy maszyna może uczestniczyć w zachowaniu komunikacyjnym na tyle bogatym, aby człowiek nie potrafił jej niezawodnie odróżnić od innego człowieka. W 2026 roku wiemy już, że w odpowiednio skonstruowanych krótkich interakcjach niektóre systemy mogą osiągać taki poziom imitacyjnej wiarygodności. Nowym pytaniem staje się więc nie "czy imitacja jest możliwa?", lecz "jaką strukturę poznawczą i jakie granice generalizacji ukrywa przekonująca imitacja?". Z tego punktu widzenia współczesny "test inteligencji" powinien przypominać raczej mapowanie krajobrazu niż przekraczanie jednej bramy. Powinniśmy badać profile stabilności, kruchości, transferu, samokorekty i relacji pomiędzy zachowaniem a mechanizmem. W jednej domenie system może wykazywać reprezentację strukturalną, w drugiej opierać się na heurystyce powierzchniowej, w trzeciej kompensować brak wiedzy przez narzędzia zewnętrzne. Ten heterogeniczny profil jest sam w sobie informacją o naturze systemu.

Tak rozumiane scattering experiments zmieniają również rolę błędu. W klasycznym benchmarku błąd obniża wynik. W eksperymencie poznawczym błąd może być materiałem diagnostycznym. Interesuje nas, czy jest przypadkowy, systematyczny, powierzchniowy, wynikający z konfliktu

reprezentacji, zbyt wczesnego zamrożenia hipotezy, błędnego samomonitorowania albo utraty reguły kontrfaktycznej.

Od autonomii procesu do odpowiedzialności normatywnej

Dwa modele o tej samej dokładności mogą posiadać zupełnie inne architektury błędu, a co za tym idzie – różną wartość poznawczą i praktyczną. Tutaj powraca podstawowa intuicja Hofstadtera: **inteligencja ujawnia się w sposobie poruszania się po przestrzeni możliwości, a nie tylko w miejscu, w którym system ostatecznie ląduje**. Idea ta, rozwijana w mikrodomenach FARG, staje się szczególnie aktualna w obliczu modeli, których poprawne odpowiedzi bywają znacznie bardziej przekonujące niż nasze rozumienie mechanizmów je wytwarzających. Zakończenie epoki Testu Turinga jako symbolicznego "egzaminu końcowego" nie oznacza zatem odejścia od koncepcji Turinga.

Jest to raczej konsekwentne rozwinięcie jego metody. Turing zastąpił pytanie metafizyczne problemem empirycznym. Dziś, gdy ta próba stała się dla wielu systemów osiągalna, powinniśmy wykonać kolejny krok: zastąpić ogólne pytanie o inteligencję AI zestawem precyzyjnych eksperymentów dotyczących reprezentacji, generalizacji, samokontroli i przyczynowej organizacji procesów. W tym punkcie Test Turinga łączy się z najtrudniejszym problemem całego artykułu: **przypisaniem zasługi za inteligentne działanie**. Jeśli model przechodzi test, tworzy analogię czy odkrywa nowe rozwiązanie, pojawia się pytanie, gdzie kończy się wkład danych, twórców architektury i użytkownika, a zaczyna autonomiczna organizacja samego procesu obliczeniowego. Problem ten, obecny już w krytyce dawnych programów przez FARG, w epoce modeli generatywnych zyskuje na złożoności. Kolejna część analizy dotyczyć będzie zatem **credit assignment**, genezy kompetencji, autorstwa oraz różnicy między przyczynową autonomią systemu a odpowiedzialnością moralną podmiotu. Komu przypisać odkrycie?

Problem **credit assignment** stanowi jeden z najgłębszych wątków programu FARG, gdyż wymusza postawienie pytania bardziej wymagającego niż to o poprawność wyniku. Jeśli system wytwarza rozwiązanie, którego programista nie przewidział szczegółowo, należy ustalić, **gdzie w łańcuchu przyczynowym znajduje się informacja odpowiedzialna za jego powstanie**. Czy rozwiązanie zostało przygotowane wcześniej przez konstruktora w postaci prymitywów i reguł? Czy wyłoniło się z interakcji elementów architektury? A może kluczowy wkład wprowadził użytkownik, właściwie formułując problem? Być może rezultat jest przede wszystkim rekombinacją wiedzy zawartej w danych lub strukturą powstałą dopiero podczas konkretnego przebiegu

obliczenia? Autor książki nazywa to właśnie problemem przypisywania zasług i wiąże z postulatem Hofstadtera, by twórczość nie była traktowana jako nieprzenikniona "czarna skrzynka". W epilogu autor formułuje tę kwestię jeszcze ostrzej: jeśli odkrycie jest logicznie przesądzone przez dostarczone prymitywy, trudno przypisać autonomiczną zasługę programowi; znacznie ciekawszy przypadek pojawia się wtedy, gdy rezultat wyłania się ze statystycznej interakcji procesów subpoznawczych i nie został jawnie zaprogramowany.

Tezę tę należy jednak doprecyzować. Fakt, że wynik nie został wpisany literalnie do programu, nie wystarcza do ustanowienia silnego sprawstwa, podobnie jak sama emergencja nie tworzy podmiotowości. Pozwala ona jednak odróżnić system, który jedynie odtwarza instrukcję, od systemu, w którym rozwiązanie zależy od wewnętrznego procesu selekcji i reorganizacji, niedookreślonego na poziomie pojedynczej reguły. Prowadzi to do kluczowego wniosku: **przyczynowa autonomia nie jest tym samym co autonomia normatywna**. System może generować zachowania niezaplanowane krok po kroku przez człowieka i w tym ograniczonym sensie posiadać znaczną autonomię procesu, ale nie czyni go to moralnym ani prawnym podmiotem odpowiedzialności. Odpowiedzialność normatywna wymaga bowiem znacznie bogatszych przesłanek: rozumienia norm, rozpoznawania powodów działania, odnoszenia ich do własnych decyzji oraz funkcjonowania w społecznych praktykach przypisywania winy i zasługi, a w niektórych teoriach – określonej formy podmiotowego doświadczenia.

FARG interesuje przede wszystkim geneza poznawcza rezultatu, którą współczesna debata musi oddzielić od problemu odpowiedzialności. Pojęcie "autorstwa AI" można zatem rozpatrywać na kilku poziomach. Pierwszym jest **autorstwo genezy kompetencji**: skąd system uzyskał zdolność wykonania zadania. Drugim – **autorstwo inicjacji**: kto zdecydował o podjęciu konkretnego działania. Trzecim jest **autorstwo organizacji lokalnego rozwiązania**: gdzie powstała struktura prowadząca do rezultatu. Czwarta warstwa dotyczy **selekcji**: kto uznał, że spośród wielu wariantów właśnie ten zostanie zachowany lub opublikowany. Piątą jest **odpowiedzialność za konsekwencje**.

W prostym narzędziu wszystkie te poziomy skupiają się po stronie człowieka. W złożonym systemie generatywnym rozkładają się one pomiędzy wielu uczestników procesu. Hofstadterowski problem okazuje się więc wczesną wersją dzisiejszego pytania o rozproszone autorstwo, czego szczególnie czystym przykładem jest Letter Spirit.

Lokalna autonomia jako wynik organizacji procesów

Projektant tworzy architekturę oraz repertuar mechanizmów. Litery wzorcowe dostarczają informacji o stylu, Cat-Evaluator stara się abstrahować cechy wspólne, Generator proponuje kolejne realizacje, Evaluator bada ich tożsamość, a Adjudicator umożliwia reorganizację interpretacji ról. Nie można zatem wskazać jednego punktu w systemie i stwierdzić, że to właśnie tam „znajduje się kreatywność”. Wynik jest efektem pętli zależności między percepcją, generowaniem, oceną a zmianą reprezentacji.

Taki model różni się zasadniczo od klasycznej maszyny kombinatorycznej. Jeśli system jedynie enumeruje wszystkie dopuszczalne warianty, a człowiek wybiera najciekawszy z nich, kluczowy element aktu selekcyjnego pozostaje poza systemem. Jeżeli natomiast mechanizmy wewnętrzne wpływają na to, które warianty są generowane, rozwijane lub odrzucane oraz jak zmieniają się kryteria w trakcie procesu, część organizacji twórczej zostaje przeniesiona do wnętrza maszyny. Autor książki podkreśla właśnie sprzężenie top-down i bottom-up w Letter Spirit: kreatywność ma wynikać nie z masowej produkcji przypadków, lecz z cyklu percepcji, ekstrakcji stylu, generowania i ewaluacji. Nie należy jednak utożsamiać tej wewnętrzności z pełną niezależnością. Letter Spirit korzysta z pojęć, ról i przestrzeni możliwości skonstruowanych przez człowieka. Jego autonomia jest więc lokalna i proceduralna, a nie absolutna.

Jest to dobra okazja, aby odejść od binarnego myślenia o sprawstwie. Pytanie „czy maszyna jest autonomiczna?” jest zbyt ogólne. Znacznie bardziej informacyjne jest pytanie: względem czego, w jakim zakresie i na którym poziomie procesu potrafi podejmować decyzje bez bezpośredniego określenia ich przez człowieka. Ten sam problem dotyczy człowieka, choć oczywiście w innej skali i strukturze.

Żaden twórca nie rozpoczyna działalności od epistemicznej próżni. Język, pojęcia, techniki, gatunki artystyczne, wiedza, szkolenie i kultura są wcześniejszymi strukturami ograniczającymi przestrzeń możliwych działań. Autor powieści nie stworzył języka, którym pisze; matematyk nie wymyślił wszystkich pojęć, których używa; kompozytor korzysta z historii harmonii, instrumentów i konwencji. Mimo to nie uznajemy ich dzieł za pozbawione autorstwa tylko dlatego, że zostały umożliwiające przez cudzy dorobek. O autorstwie przesądza nie brak zależności, lecz sposób twórczego organizowania odziedziczonego materiału.

To porównanie jest użyteczne, ale nie powinno prowadzić do pochoptnej symetrii między człowiekiem a modelem. Ludzki twórca posiada własną historię biograficzną, trwałe cele, doświadczenie świata, ciało, relacje społeczne i zainteresowania, a także możliwość odrzucenia samego zadania oraz wieloletnią ciągłość praktyki. Model generatywny może posiadać ogromną

kompetencję transformacyjną, lecz jego lokalne działanie pozostaje uruchamiane i ograniczane przez architekturę, dostępny kontekst oraz system celów zdefiniowany zewnętrznie.

Rozdzielenie odpowiedzialności przyczynowej od normatywnej

Podobieństwo dotyczy więc rozproszonej genezy twórczości, a nie pełnej równoważności podmiotów. W epoce wielkich modeli językowych problem **credit assignment** staje się jeszcze bardziej złożony, ponieważ źródła kompetencji rozkładają się na kilka warstw czasowych. Najpierw pojawia się architektura modelu, następnie ogromny zbiór danych uczących, potem proces optymalizacji, dalsze dostrajanie i post-training, instrukcje systemowe, aktualny kontekst rozmowy, ewentualna pamięć, wykorzystanie narzędzi zewnętrznych i wreszcie probabilistyczny przebieg generowania. Konkretna odpowiedź jest produktem całego tego łańcucha. Nie istnieje więc jeden prosty „autor mechanizmu”, który mógłby przewidzieć wszystkie rezultaty. W tym właśnie punkcie pojawia się nowe znaczenie autonomii generatywnej. Programista klasycznego systemu często potrafi wskazać regułę, która spowodowała dany wynik. Twórcy wielkiego modelu opiszą architekturę i procedurę uczenia, ale przed treningiem nie znają szczegółowej zawartości wszystkich reprezentacji, które się wykształcą. Jeszcze mniej realistyczne byłoby oczekiwanie, że znają każdą przyszłą odpowiedź. Powstaje zatem system, którego mechanizm genezy został zaprojektowany, lecz szczegółowa zawartość kompetencji jest w znacznej części wyuczona. To rozróżnienie można wyrazić poprzez przeciwstawienie projektowania odpowiedzi i projektowania procesu zdolnego do wytwarzania odpowiedzi.

W pierwszym przypadku człowiek przygotowuje rozwiązanie; w drugim określa warunki, w których rozwiązania mogą powstawać. Właśnie ten drugi model jest centralny dla uczenia maszynowego. Nie eliminuje on udziału człowieka, lecz przenosi go na inny poziom abstrakcji. Podobny mechanizm występuje w biologii. Hodowca nie projektuje bezpośrednio każdego szczegółu organizmu, lecz wpływa na warunki selekcji. Nauczyciel nie zapisuje gotowych przekonań w mózgu ucznia, lecz tworzy środowisko, w którym uczeń konstruuje własną reprezentację wiedzy.

Analogie te mają swoje ograniczenia, ale pokazują istotną różnicę między sterowaniem wynikiem a sterowaniem procesem generującym wyniki. Z punktu widzenia teorii przyczynowości można zatem stwierdzić, że współczesny system AI ma wielu kontrybutorów przyczynowych. Nie wszyscy jednak wnoszą ten sam rodzaj wkładu. Architektura określa klasę możliwych przekształceń. Dane dostarczają struktur, z których model może się uczyć. Funkcja celu kształtuje presję selekcyjną. Post-training ustawia część preferencji zachowania. Prompt określa lokalne zadanie.

Narzędzia zwiększają dostępny repertuar operacji, a sam przebieg inferencji dokonuje bieżącej integracji informacji. Dopiero analiza tego rozkładu pozwala sensownie pytać, komu przypisać konkretny element rozwiązania. Tak rozumiany **credit assignment** jest problemem kontrfaktycznym. Aby ustalić udział danego komponentu, trzeba zapytać: czy wynik nadal by powstał, gdyby tego komponentu zabrakło albo został zmieniony? Czy model rozwiązałby problem bez określonej informacji w prompcie? Czy ta kompetencja przetrwałaby po usunięciu pewnego rodzaju przykładów treningowych? Czy zmiana konkretnej reprezentacji wewnętrznej wpływa na rozwiązanie? Czy rezultat pozostaje podobny po wymianie modelu, przy zachowaniu promptu i tych samych danych zewnętrznych?

Takie pytania przesuwają debatę o autorstwie z poziomu intuicji na poziom analizy przyczynowej. W FARG analogiczne znaczenie miała przejrzystość mechanizmu. Autor książki wyraźnie wiąże problem zasług z postulatem, aby działania systemu były obserwowalne i związane z percepcją oraz procesem generowania. Sformułowanie, że program dzięki temu może „brać odpowiedzialność” za wybory estetyczne, powinno zostać jednak potraktowane jako metafora odpowiedzialności procesowej, a nie pełnej odpowiedzialności moralnej. To ważna korekta pojęciowa. System może być przyczynowo odpowiedzialny za wystąpienie określonej formy – w znaczeniu, że bez jego lokalnego procesu wyboru forma by nie powstała – a jednocześnie nie być podmiotem odpowiedzialności moralnej. Warto zatem rozróżnić odpowiedzialność przyczynową, odpowiedzialność funkcjonalną oraz odpowiedzialność normatywną. Pierwsza odpowiada na pytanie, co spowodowało zdarzenie. Druga wskazuje, który komponent systemu miał określone zadanie kontrolne, na przykład walidację wyniku. Trzecia odpowiada na pytanie, komu możemy zasadnie przypisać obowiązek, winę, zasługę lub sankcję. Zacieranie tych poziomów prowadzi do głębokich nieporozumień.

Rozróżnienie między autonomią funkcjonalną a odpowiedzialnością normatywną

Termostat może być przyczynowo odpowiedzialny za włączenie ogrzewania, lecz nie ponosi moralnej odpowiedzialności za temperaturę w domu. Złożony agent AI może wykazywać nieporównanie większą autonomię funkcjonalną niż termostat, jednak samo zwiększenie złożoności nie wyznacza momentu, w którym pojawia się odpowiedzialność normatywna. Problem staje się bardziej złożony, gdy system zyskuje pamięć, zdolność do samomonitorowania i korygowania własnych działań. Metacat został zaprojektowany właśnie po to, aby historia wcześniejszych prób, porażek i zmian reprezentacji mogła wpływać na przyszłe decyzje. Autor książki podkreśla, że

Copycat nie posiadał pamięci sukcesów i niepowodzeń, reprezentacji presji ani śladu temporalnego; Metacat natomiast ma uczynić część własnego procesu przedmiotem obserwacji.

Takie funkcje rozszerzają zakres autonomii epistemicznej. System jedynie reagujący na dane wejściowe ma mniejszą zdolność samodzielnego sterowania trajektorią poznawczą niż system, który potrafi funkcjonalnie stwierdzić: „ta strategia zawodziła już wcześniej, więc zmienię podejście”. Nie oznacza to jeszcze wolnej woli ani odpowiedzialności moralnej, lecz wskazuje na zmianę architektury przyczynowej: coraz więcej decyzji pośrednich zapada na podstawie reprezentacji wytworzonych przez sam system podczas wcześniejszego działania. Możemy zatem mówić o stopniach samoorganizacji procesu poznawczego. Na najniższym poziomie program wykonuje ściśle określoną sekwencję; na wyższym wybiera spośród kilku strategii w zależności od danych. Dalej może tworzyć nowe kombinacje strategii, aktualizować preferencje na podstawie historii i wreszcie budować modele własnych sukcesów oraz porażek.

Każdy z tych kroków zwiększa odległość między decyzją projektanta a szczegółem lokalnego działania. To właśnie ta odległość bywa błędnie utożsamiana z pełnym sprawstwem, podczas gdy w rzeczywistości jest to delegacja organizacji. Projektant nie kontroluje już każdego ruchu, lecz nadal określa granice systemu, zasoby, cele, dopuszczalne działania i mechanizmy ewaluacji. W architekturach agentowych część tych granic może być bardzo szeroka, ale nie znika całkowicie. Autonomia systemu oznacza zatem możliwość podejmowania decyzji wewnątrz przestrzeni działania, a nie metafizyczną niezależność od warunków, które ją ustanowiły. Takie rozumienie jest bliższe analizie autonomii organizmów i instytucji. Człowiek również funkcjonuje w granicach biologii, kultury i praw fizyki, a mimo to posiada znaczącą autonomię działania. Organizacja działa dzięki ludziom, procedurom i prawu, a jednak przypisujemy jej pewną funkcjonalną podmiotowość. Nie ma więc logicznej sprzeczności między zależnością a autonomią. Kluczowe pytanie dotyczy poziomu, na którym system potrafi samodzielnie organizować swoje działania oraz reprezentować powody ich zmiany. W przypadku AI problem komplikuje dodatkowo rola użytkownika – prompt nie jest neutralnym wyzwaczem, gdyż może zawierać znaczną część pomysłu, struktury, ograniczeń i kryteriów rezultatu.

W jednym przypadku użytkownik prosi: „napisz opowiadanie o samotności”. W drugim dostarcza fabułę, bohaterów, styl, konstrukcję scen i kluczowe metafory, pozostawiając modelowi jedynie redakcję. Oba rezultaty mogą być równie płynne językowo, lecz rozkład zasługi twórczej jest inny. Nie można zatem przypisywać autorstwa na podstawie samego rozmiaru wygenerowanego tekstu. Model może napisać tysiące słów, realizując w dużej mierze szczegółowy projekt użytkownika, albo otrzymać bardzo krótki bodziec i samodzielnie zorganizować znaczną część struktury lokalnego rezultatu. Ilość wyprodukowanych znaków nie jest miarą wkładu konceptualnego; znacznie

ważniejsze jest ustalenie, skąd pochodziły decyzje dotyczące treści, struktury, selekcji i kryterium jakości. Ta sama zasada obowiązuje wobec narzędzi.

Sprawstwo jako funkcjonalna organizacja przestrzeni możliwości

Gdy agent AI wyszukuje informacje, korzysta z kalkulatora, uruchamia kod, odwołuje się do pamięci lub przetwarza dokumenty, produkt końcowy staje się rezultatem układu hybrydowego. Nie ma powodu zakładać, że wszystkie zasługi należy przypisać komponentowi językowemu tylko dlatego, że to on formułuje finalną wypowiedź. W poznaniu człowieka również wykorzystujemy książki, notatki, kalkulatory, wyszukiwarki i innych ludzi; granice indywidualnej inteligencji od dawna są częściowo rozszerzane przez środowisko poznawcze. Prowadzi to do teorii **poznania rozproszonego**. Zadanie może zostać rozwiązane przez układ obejmujący człowieka, model, narzędzie i zewnętrzną pamięć, przy czym żadna z tych części samodzielnie nie posiada pełnej struktury rozwiązania.

W takim układzie pytanie „kto to zrobił?” nie zawsze znajduje jedną, prostą odpowiedź. Można nadal wyznaczać odpowiedzialność społeczną i instytucjonalną, ale wyjaśnienie poznawcze musi uwzględniać system jako całość. Hofstadterowskie **credit assignment** zyskuje tutaj nowe znaczenie. Nie dotyczy już tylko tego, czy programista „przemycił” rozwiązanie do systemu, lecz odnosi się do dekompozycji całego aktu poznawczego. Które elementy były niezbędne, a które zastępowalne? Które wniosły strukturę, a które jedynie umożliwiły jej realizację? Gdzie pojawiła się nowa informacja? Który komponent rozpoznał, że wcześniejsza strategia jest niewystarczająca i kto lub co dokonało selekcji końcowego wariantu?

To ostatnie pytanie jest szczególnie istotne. W historii sztuki i nauki generowanie możliwości nigdy nie stanowiło całego procesu twórczego; kluczową rolę odgrywa selekcja. Pisarz usuwa większość pierwszych wersji, naukowiec porzuca hipotezy, projektant odrzuca szkice. Jeżeli system sztucznej inteligencji jedynie produkuje warianty, a człowiek dokonuje całej jakościowej selekcji, ludzki wkład pozostaje centralny. Jeśli jednak system sam hierarchizuje propozycje, porównuje je z kryteriami i wraca do generowania po rozpoznaniu słabości, jego udział poznawczy rośnie. Model Letter Spirit jest pod tym względem istotny, gdyż generator nie działa w nim w izolacji.

Evaluator, Adjudicator i Cat-Evaluator tworzą warstwę selekcji i reinterpretacji. Twórczość zostaje zlokalizowana nie w jednym module generującym, lecz w pętli, w której forma staje się przedmiotem oceny, a ta z kolei wpływa na kolejną generację. Rozwiązanie to jest konceptualnie

bliskie dzisiejszym systemom iteracyjnym, w których model może wygenerować odpowiedź, poddać ją krytyce, skorzystać z narzędzia weryfikującego i stworzyć poprawioną wersję. Nie należy jednak automatycznie uznawać takiej pętli za „samokrytykę” w ludzkim sensie. Jej wartość zależy od niezależności i jakości mechanizmu oceny. Jeżeli ten sam błąd występuje zarówno w generatorze, jak i w krytyku, iteracja może jedynie wzmacniać złudzenie poprawności. Autonomia procesu bez epistemicznej różnorodności może prowadzić do autokonfirmacji. Metacatowska intuicja o wykrywaniu własnego zapętlenia jest więc równie ważna, co sama zdolność do ponownego generowania.

W tym miejscu pojawia się możliwość sformułowania bardziej precyzyjnego kryterium maszynowego sprawstwa. Nie należałoby pytać wyłącznie, czy rezultat był nieprzewidziany, gdyż losowy generator również wytwarza rezultaty nieprzewidziane. Znacznie bardziej istotne jest to, czy system **sam organizuje przestrzeń możliwości według reprezentowanych kryteriów, wykorzystuje historię własnych działań, wykrywa konflikty, rekonfiguruje strategię i selekcionuje rezultat w sposób przyczynowo zależny od tych wewnętrznych stanów.** Tak rozumiane sprawstwo jest funkcjonalne i stopniowalne. Nie wymaga przypisywania maszynie intencji fenomenalnej, pozwala jednak odróżnić nieprzewidywalność czysto losową od tej wynikającej ze złożonej organizacji poznawczej. Ta różnica jest zasadnicza: rzut kostką może zaskoczyć człowieka, ale nie interpretuje wyniku.

Rozdzielenie sprawstwa technicznego od odpowiedzialności normatywnej

System twórczy powinien wykorzystywać własny wynik jako informację modyfikującą kolejne działania. Prowadzi to do ponownego zdefiniowania emergencji: emergentny rezultat nie jest wartościowy dlatego, że projektant nie potrafił go przewidzieć. Nieprzewidywalność może bowiem wynikać z chaosu, złożoności lub zwykłego braku informacji. Poznawczo interesująca emergencja występuje wtedy, gdy z lokalnych interakcji wyłania się stabilna organizacja posiadająca funkcjonalną moc regulacyjną.

W Copycat byłaby nią koherentna interpretacja analogii, w Letter Spirit – styl, a w Metacat – reprezentacja własnej trajektorii poznawczej. Sama niespodzianka nie jest jeszcze inteligencją, co ma kluczowe znaczenie dla współczesnego dyskursu. Często twierdzi się, że system „zaskoczył swoich twórców”, jakby zaskoczenie automatycznie dowodziło autonomii. Tymczasem można zaskoczyć twórcę wynikiem błędnym, przypadkowym lub wynikającym z nieprzejrzystej kombinatoryki. Silniejszym dowodem jest powstanie kompetencji systematycznej, transferowalnej,

podatnej na regulację i wykorzystywanej przez sam system w nowych warunkach. Z tego względu credit assignment powinien obejmować również testy ablacyjne i kontrfaktyczne. Jeśli usunięcie określonego komponentu nie zmienia rezultatu, jego udział był mniejszy, niż sugerowała intuicja. Jeśli zaś po zmianie kilku przykładów treningowych znika konkretna kompetencja, można zacząć lokalizować jej genezę.

Jeśli odpowiedź modelu radykalnie zależy od drobnego sformułowania promptu, udział użytkownika i kontekstu powierzchniowego jest większy. Jeżeli natomiast struktura pozostaje stabilna mimo wymiany promptu, danych czy narzędzi, istnieją silniejsze podstawy do lokalizowania kompetencji w samym modelu. Credit assignment jest zatem problemem empirycznym, a nie wyłącznie filozoficznym; można projektować eksperymenty przyczynowe, które stopniowo rozdzielają wkłady poszczególnych elementów. Oczywiście nigdy nie otrzymamy absolutnie czystego rozkładu zasługi, ponieważ systemy złożone wykazują interakcje – efekt jednego komponentu zależy od obecności innych. Niemniej można budować coraz lepsze modele zależności.

W wymiarze społecznym konieczny jest jednak drugi krok. Z faktu, że maszyna była istotną przyczyną rezultatu, nie wynika, że społeczeństwo powinno przenieść na nią odpowiedzialność. W praktykach odpowiedzialności nie chodzi bowiem tylko o przyczynę, lecz także o kontrolę, przewidywalność, obowiązek nadzoru i możliwość reagowania na normy. Projektant, operator, organizacja lub użytkownik mogą pozostawać odpowiedzialni właśnie dlatego, że zdecydowali o wdrożeniu systemu, określili jego zakres działania albo mieli możliwość ograniczenia ryzyka.

Pojawia się tutaj istotny paradoks automatyzacji: im bardziej autonomicznie system wykonuje zadanie, tym łatwiej człowiekowi psychologicznie dystansować się od rezultatu. Jednocześnie im mniej przejrzysty jest mechanizm, tym trudniej przypisać konkretny błąd pojedynczej osobie. Powstaje luka odpowiedzialności, lecz nie należy jej rozwiązywać poprzez automatyczne personifikowanie maszyny. Odpowiedzialność należy raczej projektować na poziomie całego systemu społeczno-technicznego. W tym miejscu problem credit assignment łączy się z architekturą aksjologii zapowiedzianą w temacie książki. Skoro przyszłych możliwości technologicznych nie potrafimy w pełni przewidzieć, potrzebujemy względnie stabilnych kryteriów oceny sposobu delegowania działania. Nie możemy opierać etyki wyłącznie na tym, czy system zachowuje się „inteligentnie”. Potrzebujemy pytań o możliwość kontroli, identyfikowalność decyzji, zgodność z prawdą, rozkład ryzyka, ochronę autonomii człowieka oraz możliwość przypisania odpowiedzialności za konsekwencje.

Konieczność teorii procesu dla przypisania zasługi i odpowiedzialności

Płynność inteligencji nie może oznaczać płynności odpowiedzialności aż do jej całkowitego rozpuszczenia. Wręcz przeciwnie: im bardziej rozproszona staje się przyczynowa organizacja działania, tym dokładniej należy rekonstruować jej węzły. Inteligentny system może być wieloskładnikowy, emergentny i częściowo nieprzewidywalny, lecz instytucjonalne reguły odpowiedzialności powinny pozostać na tyle stabilne, by jasno określać, kto decyduje o dopuszczeniu systemu do użytku, kto wyznacza granice jego autonomii i kto ponosi odpowiedzialność za szkodliwe skutki. Architektura aksjologiczna pełni tu rolę podobną do głębokich inwariantów w programie Copycat.

Powierzchniowe elementy technologii – modele, interfejsy, metody treningu czy zakres autonomii – mogą zmieniać się gwałtownie. Nie oznacza to jednak, że każda zmiana techniczna wymaga budowania moralności od podstaw. Pewne wartości: prawda, odpowiedzialność, autonomia, godność, proporcjonalność czy możliwość zakwestionowania decyzji, mogą pełnić funkcję stabilnych punktów odniesienia, względem których oceniamy kolejne konfiguracje technologiczne. Choć analogia ta ma swoje granice, pozostaje intelektualnie użyteczna. W płynnej inteligencji nie wszystko musi być płynne. Kreatywna transformacja wymaga inwariantów pozwalających ocenić, czy struktura zachowała sens. Podobnie innowacja technologiczna potrzebuje kryteriów normatywnych, by rozstrzygnąć, czy zmiana nadal służy człowiekowi, czy też rozpuszcza odpowiedzialność w anonimowej sieci procesów.

Problem zasługi prowadzi zatem poza pytanie o to, czy maszyna jest „autorem”. Bardziej fundamentalne staje się pytanie o strukturę sprawstwa w układzie człowiek-model-dane-narzędzia-instytucja. Można w nim wyróżnić wkład koncepcyjny, proceduralny, selekcyjny, wykonawczy i normatywny. Role te mogą być rozdzielone między różne elementy systemu; dopiero ich precyzyjne wyodrębnienie pozwala sensownie mówić o autorstwie. Z perspektywy naukowej kluczowe jest uniknięcie dwóch skrajności.

Pierwsza zakłada, że „maszyna niczego nie robi, bo wszystko pochodzi od programistów i danych”. Jest to teza zbyt słaba, gdyż ignoruje autonomię lokalnego procesu i kompetencje emergentne, których nikt nie zaprogramował wprost. Druga twierdzi, że „maszyna sama stworzyła rezultat, ponieważ nikt nie przewidział jego dokładnej postaci”. Jest ona równie niewystarczająca, bo pomija genezę architektury, dobór danych, cele i samą inicjację zadania. Trafniejszy obraz przedstawia twórczość i inteligencję jako łańcuch transformacji, w którym informacja zmienia poziom organizacji: ludzka kultura dostarcza danych, projektanci tworzą mechanizm zdolny do uczenia,

trening kompresuje regularności w parametrach, kontekst aktywuje konkretne kompetencje, a obliczenie konstruuje odpowiedź. Na każdym z tych poziomów pojawia się inny rodzaj wkładu i inny sens „zasługi”.

Właśnie dlatego pytanie „czy AI stworzyła ten tekst?” jest zazwyczaj zbyt ubogie. Znacznie głębsze pytania brzmią: jaka część idei została zadana przez człowieka, co pochodzi z korpusu kulturowego, jakie relacje model utworzył lokalnie, kto dokonał selekcji, kto zweryfikował prawdziwość i kto zdecydował o publikacji. Dopiero taki opis oddaje epistemiczną strukturę współczesnej współpracy człowieka z maszyną. Hofstadterowski credit assignment okazuje się więc nie historyczną ciekawostką, lecz problemem, który urósł wraz z sukcesem systemów generatywnych. W mikrodomenie Letter Spirit chodziło o ustalenie, czy elegancja formy wynika z rzeczywistej percepcji i reinterpretacji, czy została ukryta w ręcznie zaprojektowanych regułach. Dziś pytamy analogicznie: czy abstrakcja modelu wynika z wyuczonej struktury, czy z pamięci danych; czy samokorekta jest stabilnym mechanizmem, czy efektem promptu; czy kreatywność obejmuje selekcję, czy jedynie produkcję wariantów.

W każdym z tych przypadków odpowiedź wymaga obserwacji procesu. I tutaj domyka się jedna z najważniejszych nici programu FARG: bez teorii procesu nie można wiarygodnie przypisać zasługi wynikowi. Produkt może być imponujący, lecz jego epistemiczne znaczenie pozostaje niejasne, dopóki nie znamy struktur, które doprowadziły do jego powstania. To stwierdzenie przygotowuje grunt pod kolejny etap. Skoro sprawstwo i autorstwo są stopniowalne oraz rozproszone, należy zapytać, jakie wartości powinny pozostać stabilne w obliczu technologii, która coraz bardziej zaciera granice między twórcą, narzędziem, współpracownikiem a autonomicznym procesem. Wówczas analiza przejdzie od epistemologii inteligencji do architektury aksjologii: prawdy, odpowiedzialności, autonomii, godności i kontroli jako normatywnych inwariantów w świecie, którego technicznego kształtu nie potrafimy wiarygodnie przewidzieć.

W świecie, gdzie granica między twórcą a narzędziem ulega rozmyciu, najwyższym ryzykiem nie jest stworzenie maszyny zbyt inteligentnej, lecz dopuszczenie do sytuacji, w której płynność technologii całkowicie rozpuszcza stabilność odpowiedzialności. Czy potrafimy zaprojektować systemy etyczne, które pozostaną niezmiennymi inwariantami w obliczu procesów poznawczych, których kształtu nie jesteśmy już w stanie przewidzieć? Ostatecznie to nie zdolność maszyny do myślenia, lecz nasza zdolność do definiowania granic sprawstwa zdecyduje o tym, czy pozostaniemy podmiotami, czy jedynie administratorami rozproszonych procesów.

Inspiracje

[1]



"Fluid Concepts and Creative Analogies" Douglas Hofstadter

ISBN: 9780140258356

Douglas Richard Hofstadter (ur. 15 lutego 1945 r.) jest amerykańskim uczonym i wybitnym profesorem kognitywistyki i literatury porównawczej na Uniwersytecie Indiany w Bloomington. Jest znany na całym świecie ze swoich pionierskich, interdyscyplinarnych badań łączących kognitywistykę, informatykę, matematykę, psychologię i filozofię umysłu. Hofstadter zyskał światowe uznanie za swoje arcydzieło z 1979 r. „Gödel, Escher, Bach: wieczny złoty warkocz”, które otrzymało Nagrodę Pulitzera w kategorii literatury faktu oraz National Book Award. Jego prace intelektualne koncentrują się głównie na mechanizmach ludzkiego myślenia, koncentrując się na koncepcjach dziwnych pętli, autoreferencyjności, świadomości i centralnej roli tworzenia analogii w twórczym poznaniu. Poprzez grupę badawczą Fluid Analogies Research Group kierował rozwojem architektur obliczeniowych modelujących emergentne procesy umysłowe, metafory i poślizgi konceptualne, pozostawiając głębokie i trwałe dziedzictwo zarówno w dziedzinie sztucznej inteligencji, jak i filozofii kognitywnej.

Douglas Richard Hofstadter (born February 15, 1945) is an American scholar and Distinguished Professor of Cognitive Science and Comparative Literature at Indiana University Bloomington. He is internationally recognized for his pioneering multidisciplinary research bridging cognitive science, computer science, mathematics, psychology, and philosophy of mind. Hofstadter gained worldwide acclaim for his 1979 masterpiece Gödel, Escher, Bach: an Eternal Golden Braid, which was awarded the Pulitzer Prize for General Nonfiction and the National Book Award. His intellectual work primarily investigates the mechanisms of human thought, focusing on the concepts of strange loops, self-reference, consciousness, and the central role of analogy-making in creative cognition. Through the Fluid Analogies Research Group, he guided the development of computational architectures modeling emergent mental processes, metaphor, and conceptual slippage, leaving a profound and lasting legacy across both artificial intelligence and cognitive philosophy.

Indiana University Bloomington

Literatura uzupełniająca

Książki

Książki autora i dzieła przywołane w tekście:



[1] "The Mind's I"

Douglas Hofstadter

Hachette UK | ISBN: 9780786725144



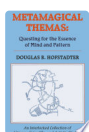
[2] "Gödel, Escher, Bach"

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

Prace autorów przywoływanych w tekście:

[1] "Computing Machinery and Intelligence (1950)"

Alan Turing

2004 | The Essential Turing | DOI: 10.1093/oso/9780198250791.003.0017

[Źródło](#)

[2] "ALAN M. TURING'S CRITIQUE OF RUNNING SHORT CRIBS ON THE U. S. NAVY BOMBE"

ALAN M. TURING

2003 | Cryptologia | DOI: 10.1080/0161-110391891748

[Źródło](#)

[3] "Dusting Off the Turing Test"

A. Hodges, A. Turing

2012

[Źródło](#)

[4] "Dystopian or Utopian? Two Images of Alan Turing"

A. Turing

2022

[Źródło](#)

Artykuły naukowe autora:

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

Jonathan Lieberman, Douglas R. Hofstadter

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

[Źródło](#)

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

Douglas R. Hofstadter

1995

[Źródło](#)

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

Charles Taylor, Douglas R. Hofstadter

1985 | The Antioch Review | DOI: 10.2307/4611504

[Źródło](#)

[8] "Fluid Concepts and Creative Analogies"

Douglas R. Hofstadter

1995

[Źródło](#)

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

Douglas R. Hofstadter

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

[Źródło](#)

[10] "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](#)

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

Douglas R. Hofstadter, Melanie Mitchell

1995

[Źródło](#)

[12] "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: cb2075f1-9786-4b57-b4b8-cfb99d14ba62