

Płynność inteligencji: od analogii Douglasa Hofstadtera do współczesnych modeli językowych



AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Tekst analizuje naturę inteligencji nie jako sprawne operowanie na stałych symbolach, lecz jako zdolność do ciągłej rekonfiguracji reprezentacji problemu. Przywołując program badawczy Fluid Analogies Research Group (FARG) Douglasa Hofstadtera, autor wskazuje na przejście od sztywnych algorytmów ku „wierności kognitywnej”. Kluczowym elementem jest tu płynność pojęciowa (conceptual slippage), która pozwala systemowi poznawczemu dostosowywać granice kategorii w celu uchwycenia głębszych struktur i regularności. W przeciwieństwie do klasycznej AI, gdzie reprezentacja jest dana z góry, podejście to zakłada sprzężenie zwrotne między rozumowaniem a percepcją, co stanowi fundament dla zrozumienia współczesnych modeli językowych i ich zdolności do elastycznego przetwarzania informacji.

The text analyzes the nature of intelligence not as the efficient manipulation of fixed symbols, but as the ability to continuously reconfigure the representation of a problem. Drawing on Douglas Hofstadter's Fluid Analogies Research Group (FARG) research program, the author points to a transition from rigid algorithms toward "cognitive fidelity." A key element here is conceptual slippage, which allows a cognitive system to adjust category boundaries in order to capture deeper structures and regularities. In contrast to classical AI, where representation is predefined, this approach assumes a feedback loop between reasoning and perception, providing the foundation for understanding modern language models and their capacity for flexible information processing.

Artykuł analizuje naturę inteligencji przez pryzmat teorii płynności poznawczej Douglasa Hofstadtera i programu badawczego FARG. Główna teza tekstu głosi, że istotą

myślenia nie jest sprawne operowanie wewnątrz raz zadanej struktury danych, lecz zdolność do dynamicznej rekonfiguracji samych reprezentacji świata – czyli umiejętność zmiany sposobu, w jaki problem zostaje rozpoznany i sformułowany. Autor argumentuje, że inteligencja przejawia się w balansie między stabilnością a elastycznością (tzw. poślizgiem pojęciowym), co pozwala na tworzenie głębokich analogii poprzez wybór odpowiednich inwariantów strukturalnych zamiast polegania na powierzchownym podobieństwie. Analizując modele takie jak Copycat i Seek-Whence, tekst ukazuje, że procesy percepcji i rozumowania są ze sobą nierozzerwalnie sprzężone: myślenie jest w istocie formą aktywnego konstruowania reprezentacji. W kontekście współczesnych modeli językowych (LLM), artykuł stawia prowokacyjne pytanie o różnicę między behawioralną kompetencją a rzeczywistym mechanizmem rekonfiguracji pojęciowej, sugerując, że tradycja FARG dostarcza kluczowych narzędzi diagnostycznych do badania tego, czy sztuczny umysł rzeczywiście rozumie relacje, czy jedynie sprawnie nawiguje po statystycznym krajobrazie form.

SŁOWA KLUCZOWE

płynność inteligencji

Douglas Hofstadter

modele językowe

rekonfiguracja reprezentacji

Fluid Analogies Research Group

emergencja w AI

conceptual slippage

kodlety

równoległy skan tarasowy

inwarianty strukturalne

przestrzeń embeddingów

metaracjonalność

testy kontrfaktyczne

płynność pojęciowa

kognitywna wierność

Inteligencja jako zdolność do rekonfiguracji reprezentacji

Płynność inteligencji. Od analogii Hofstadtera do modeli językowych: percepcja, kreatywność, samoobserwacja i granice przewidywania sztucznego umysłu

Historię sztucznej inteligencji można odczytywać jako kronikę kolejnych odpowiedzi na pytanie o naturę myślenia. Nie jest to jedynie spór technologiczny o skuteczniejsze algorytmy, większe pamięci, szybsze procesory czy obszerniejsze zbiory danych. Pod powierzchnią inżynierskich debat trwa znacznie starsza dyskusja filozoficzna: czy inteligencja polega przede wszystkim na

poprawnym wykonywaniu operacji na symbolach, czy raczej na zdolności do ciągłego przekształcania sposobu, w jaki sytuacja zostaje przez podmiot rozpoznana i przedstawiona. W pierwszym ujęciu struktura problemu jest w zasadniczym stopniu dana z góry, a inteligencja wykonuje na niej operacje. W drugim zaś sama struktura problemu staje się przedmiotem poznawczej pracy. Nie wystarczy wtedy poprawnie manipulować reprezentacją; trzeba umieć wytworzyć reprezentację właściwą, a gdy ta przestaje wyjaśniać sytuację – przebudować ją. W tym właśnie przesunięciu, od operowania pojęciami do operowania również granicami pojęć, kryje się jeden z najważniejszych intelektualnych wkładów tradycji Douglasa Hofstadtera i Fluid Analogies Research Group.

Program badawczy FARG wyrósł z nieufności wobec obrazu inteligencji, który w klasycznej sztucznej inteligencji coraz częściej utożsamiał poznanie z formalnym przetwarzaniem jawnie określonych symboli. Hofstadter trafnie ujmuje tę zmianę jako odejście od sztywnej reprezentacji ku „wierności kognitywnej”: zamiast pytać przede wszystkim, czy program rozwiązuje zadanie, on i jego współpracownicy pytali, czy mechanizm rozwiązania przypomina procesy, dzięki którym człowiek orientuje się w strukturze problemu. Z tego powodu FARG skoncentrowała się na starannie dobranych mikrodomenach – sekwencjach liczbowych, anagramach, prostych zadaniach arytmetycznych, analogiach literowych, układach przedmiotów i stylu graficznym – traktując je nie jako zabawki obliczeniowe, lecz jako laboratoryjnie uproszczone środowiska, w których można obserwować mechanizmy reorganizacji pojęciowej.

Nie oznaczało to odrzucenia formalizacji. FARG nie była antyobliczeniową teorią umysłu. Przeciwnie, jej przedstawiciele budowali programy komputerowe, aby wykazać, że obliczenie nie musi być utożsamiane z liniowym stosowaniem reguł do raz zdefiniowanych symboli. Różnica dotyczyła raczej poziomu, na którym lokowano istotę inteligencji. W klasycznej tradycji symbolicznej reprezentacja problemu bywała przygotowana przed rozpoczęciem właściwego wnioskowania: świat dzielono na obiekty, cechy i relacje, a następnie algorytm poszukiwał rozwiązania w tak skonstruowanej przestrzeni.

Hofstadter kwestionował właśnie tę separację. Uznawał, że tworzenie reprezentacji oraz rozwiązywanie problemu są ze sobą sprzężone. To, co system postrzega jako „obiekt”, „relację”, „grupe”, „wyjątek”, „początek”, „koniec”, „następstwo” lub „podobieństwo”, nie jest neutralnym wejściem do rozumowania, lecz jego rezultatem. Rozumowanie zmienia percepcję, a zmieniona percepcja przekształca kolejne rozumowanie. Program FARG można zatem umieścić w długiej historii sporów o reprezentację.

Filozofia poznania od dawna mierzyła się z problemem, czy umysł odbija strukturę świata, czy też uczestniczy w jej kategoryzowaniu. Psychologia postaci wykazała, że percepcja nie jest biernym

sumowaniem bodźców, lecz organizowaniem pola w figury, tła, całości i relacje. Badania nad rozwiązywaniem problemów typu insight wskazywały później, że przełom intelektualny często nie polega na wykonaniu większej liczby kroków w dotychczasowej reprezentacji, lecz na zmianie samej reprezentacji zadania. Kognitywistyka drugiej połowy XX wieku odziedziczyła te intuicje i nadała im postać modeli obliczeniowych.

FARG zajęła w tym pejzażu pozycję osobliwą: nie odrzucała symboli, lecz nie ufała symbolom nieruchomym; nie odrzucała sieciowych skojarzeń, lecz nie chciała redukować semantyki do pozbawionej interpretacji aktywacji; nie odrzucała reguł, lecz sądziła, że prawdziwe poznanie wymaga również zdolności do rozpoznawania sytuacji, w których dotychczasowa reguła przestaje być właściwym opisem problemu. Centralnym pojęciem tej tradycji stała się płynność pojęciowa. Płynność nie oznacza dowolności, chaosu ani niestabilności wiedzy. Oznacza raczej zdolność systemu poznawczego do takiego dostosowania granic kategorii, aby zachować ważniejsze relacje strukturalne mimo zmiany powierzchownych cech sytuacji. Hofstadter opisuje to za pomocą pojęć conceptual halos i conceptual slippage: pojęcia posiadają otoczki możliwych znaczeń i powiązań, a pod presją kontekstu jedno rozróżnienie może zostać zastąpione innym. „Prawo” może przejść w „lewo”, „następnik” w „poprzednik”, „pierwszy” w „ostatni”, o ile takie przesunięcie pozwala uchwycić głębszą regularność. Nie jest to pomyłka logiczna, lecz kontrolowana zmiana poziomu opisu. Właśnie tutaj pojawia się słynna metafora wody.

Hofstadter wskazywał, że płynność cieczy nie jest własnością pojedynczej cząsteczki H_2O . Na poziomie mikroskopowym obserwujemy ruch, powstawanie i rozpad lokalnych konfiguracji, natomiast na poziomie makroskopowym powstaje stabilna właściwość, którą nazywamy płynnością. W modelach FARG podobną rolę pełnią liczne drobne procesy podpoznawcze: lokalne, częściowo stochastyczne, nieposiadające samodzielnie całej struktury problemu.

Emergencja jako relacja poziomów a pułapka predefiniowanych reprezentacji

Z ich współdziałania ma wyłaniać się rozwiązanie spójne na poziomie wysokim. Autor określa to podłożem podpoznawczym i wiąże z działaniem kodletów – małych agentów wykonujących lokalne czynności, których zbiorowa dynamika umożliwia powstawanie większych struktur poznawczych. Metaforę tę należy jednak interpretować rygorystycznie. Nie jest ona teorią neurobiologiczną i nie dowodzi, że mózg działa jak ciecz, że świadomość jest produktem procesów formalnie identycznych z dynamiką wiązań wodorowych ani że sama stochastyczność wystarcza do wytworzenia inteligencji. Jej wartość ma charakter metodologiczny.

Wskazuje ona na możliwość istnienia własności systemowych, które nie są jawnie zapisane w żadnym pojedynczym składniku systemu. W języku teorii systemów oznaczałoby to, że interesująca właściwość ujawnia się na innym poziomie opisu niż poziom elementarnych interakcji. Tak rozumiana emergencja nie jest tajemniczą siłą, lecz relacją między poziomami organizacji: lokalne procesy podlegają jednemu opisowi, natomiast regularność globalna wymaga opisu innego. Problem poznawczy zaczyna się dopiero wtedy, gdy próbuje się ustalić, czy tę regularność można w pełni przewidzieć, zredukować lub wyjaśnić na podstawie poziomu niższego. To rozróżnienie będzie miało zasadnicze znaczenie również dla późniejszego namysłu nad modelami językowymi. W debacie o współczesnej AI termin „emergencja” bywa używany w wielu różnych znaczeniach: jako określenie jakościowo nowej kompetencji pojawiającej się przy zwiększaniu skali modelu, opis nieliniowej zmiany wyniku benchmarku, metafora nieprzejrzystości procesu uczenia albo sugestia powstania czegoś ontologicznie nowego. Są to twierdzenia różnego rodzaju i nie należy ich mieszać.

Hofstadterowska intuicja jest tu bardziej wymagająca: autora interesuje nie samo pojawienie się zachowania, którego programista nie wpisał jawnie do kodu, lecz powstanie elastycznej organizacji reprezentacyjnej, zdolnej do reorganizacji własnego sposobu ujmowania problemu. To znacznie wyższy próg niż proste stwierdzenie, że system zrobił coś zaskakującego. Właśnie dlatego należy ostrożnie rozumieć również krytykę tradycyjnej sztucznej inteligencji obecną w książce. Historyczne programy symboliczne nie były jedną jednorodną szkołą ani prostymi maszynami wykonującymi wyłącznie „twarde reguły”. W tradycji symbolicznej rozwijano wyszukiwanie heurystyczne, planowanie, rozumowanie probabilistyczne, systemy produkcyjne, uczenie pojęć oraz reprezentacje ramowe i semantyczne. Krytyka FARG była bardziej precyzyjna niż stereotypowe przeciwstawienie „logika kontra twórczość”: jej celem było pokazanie, że jeśli reprezentacja zostanie zbyt szczegółowo ustalona przed rozpoczęciem rozumowania, wówczas najtrudniejsza część inteligencji może zostać nieświadomie wykonana przez projektanta.

Program otrzymuje wtedy świat już pokrojony na istotne elementy. Potrafi poprawnie wnioskować, ponieważ człowiek wcześniej zdecydował, co w ogóle liczy się jako właściwy obiekt wnioskowania. Z tego punktu widzenia problem przypisania zasług – **credit assignment** – pojawia się znacznie wcześniej niż w dyskusjach o generatywnej AI. Jeżeli program odkrywa wzorzec, trzeba zapytać, jaka część tego wzorca była już zakodowana w jego reprezentacji, języku prymitywów, heurystykach, funkcji oceny lub sposobie przygotowania danych wejściowych. Autor wykorzystuje tę intuicję w krytyce takich systemów jak AM, AARON czy różnych architektur opartych na wcześniej ustalonych reprezentacjach, sugerując, że obserwator może przypisać maszynie odkrycie, którego warunki zostały w rzeczywistości głęboko przygotowane przez człowieka. Problem ten nie stracił aktualności.

Przeciwnie, w epoce systemów uczonych na ogromnych korpusach danych stał się trudniejszy, ponieważ udział człowieka nie znajduje się już wyłącznie w kodzie programu, lecz rozprasza się pomiędzy architekturę, funkcję celu, selekcję danych, procedury trenowania, ocenę ludzką, instrukcje systemowe, kontekst użytkownika oraz narzędzia zewnętrzne.

Inteligencja jako dynamiczna równowaga między eksploracją a strukturą

FARG odpowiadała na ten problem poprzez próbę rozproszenia kontroli poznawczej. W modelach takich jak Copycat nie istnieje pojedynczy, centralny moduł, który posiadałby kompletne rozwiązanie i wydawał instrukcje pozostałym elementom. Kodlety badają lokalne możliwości: jedne tworzą struktury, inne je osłabiają lub niszczą, a prawdopodobieństwo ich uruchomienia zależy od aktualnego stanu całego systemu. Zachowanie ma zatem wynikać z pętli zależności między poziomem lokalnym a globalnym. Struktury wyłaniające się w przestrzeni roboczej zmieniają szanse na późniejsze działania kodletów, podczas gdy same te działania modyfikują struktury globalne. Autor określa ten mechanizm jako sprzężenie bottom-up i top-down oraz równoległy skan tarasowy: wiele hipotez może być badanych jednocześnie, lecz nie z jednakową intensywnością.

Tego rozwiązania należy interpretować jako model ograniczonej racjonalności zasobowej. System inteligentny nie jest w stanie szczegółowo zbadać wszystkich możliwych interpretacji – przestrzeń kandydatów rośnie zbyt szybko, a część z nich jest pozbawiona sensu już na podstawie słabych wskazówek kontekstowych. Inteligencja musi więc równoważyć eksplorację z eksploatacją: badać nowe możliwości, lecz jednocześnie inwestować coraz więcej zasobów tam, gdzie pojawiają się obiecujące struktury.

Równoległy skan tarasowy jest czymś więcej niż techniczną sztuczką FARG. Wyraża ogólniejszą zasadę poznawczą: sprawne rozumowanie nie polega ani na mechanicznym sprawdzeniu wszystkich wariantów, ani na natychmiastowym zamknięciu się w pierwszej hipotezie. Polega raczej na nierównomiernym i dynamicznym przydzielaniu uwagi wielu konkurencyjnym sposobom interpretacji. Kluczowy jest tutaj problem temperatury. W Copycat wysoka temperatura odpowiada sytuacji, w której zbudowane struktury są mało spójne i system powinien dopuszczać większą różnorodność działań. Niska temperatura pojawia się wówczas, gdy rozwiązanie zaczyna tworzyć zwartą strukturę, co pozwala ograniczyć eksplorację i stabilizować interpretację. Autor przedstawia temperaturę jako wewnętrzny sygnał jakości organizacji przestrzeni roboczej, a nie jedynie zewnętrznie narzucony parametr losowości. To rozróżnienie prowadzi do istotnej tezy:

stochastyczność jest użyteczna poznawczo wtedy, gdy pozostaje sprzężona z oceną struktury problemu.

Losowość sama w sobie nie tworzy inteligencji. Twórcza eksploracja wymaga mechanizmu, który wie – funkcjonalnie, niekoniecznie świadomie – kiedy należy zwiększyć tolerancję na odległe hipotezy, a kiedy przeciwnie: zacząć wzmacniać lokalną spójność. W perspektywie psychologicznej jest to bliskie obserwacji, że człowiek zmienia strategię poznawczą wraz ze zmianą poczucia postępu. Gdy rozwiązanie wydaje się bliskie, uwaga zwęża się, a kolejne operacje służą dopracowaniu istniejącej reprezentacji. Gdy pojawia się impas, niezbędna staje się większa elastyczność, poluzowanie założeń i dopuszczenie alternatywnej segmentacji problemu. W badaniach nad twórczością często mówi się o napięciu między generowaniem a oceną, rozbieżnością a selekcją, eksploracją a kontrolą. Model FARG nadał tej intuicji formę architektoniczną: system ma umieć być czasem bardziej „gorący”, a czasem bardziej „zimny”, ponieważ inteligencja wymaga zarówno tolerowania niepewności, jak i zamykania poszukiwań w momencie, gdy struktura staje się wystarczająco dobra.

Najbardziej znaczącą konsekwencją filozofii płynnych pojęć jest jednak przesunięcie granicy między percepcją a myśleniem. W potocznym obrazie percepcja dostarcza danych, a myślenie je przetwarza. FARG podważa tę kolejność. Percepcja wysokiego poziomu sama w sobie jest już rodzajem wnioskowania; system musi bowiem zdecydować, które elementy należą do wspólnej grupy, jaka relacja jest ważniejsza od innych, co stanowi tło, a co figurę oraz który wyjątek należy zignorować, a który wymaga przebudowy całej interpretacji. Autor formułuje to wprost: „myślenie jako percepcja” oraz „poznanie jako rozpoznawanie” oznaczają, że rozwiązywanie problemów polega na budowaniu coraz trafniejszych reprezentacji świata, a nie wyłącznie na stosowaniu reguł do reprezentacji raz zadanej.

Takie ujęcie prowadzi do istotnego rozróżnienia między błędem w obrębie modelu a błędem modelu. Jeżeli system posiada właściwy sposób reprezentowania sytuacji, lecz wykonuje niewłaściwą operację, jego problem jest lokalny. Jeżeli jednak sama reprezentacja jest nietrafna, nawet perfekcyjna logika może prowadzić do rezultatów poznawczo jałowych. Człowiek rozwiązuje wiele trudnych problemów właśnie poprzez zmianę tego, co uznaje za ich elementarną strukturę. Klasycznym przykładem są zadania, w których nagle rozwiązanie następuje po dostrzeżeniu nowego grupowania obiektów albo zmianie znaczenia ograniczenia. FARG uznaje ten rodzaj „widzenia jako” za centralny dla inteligencji.

Inteligencja jako zdolność rekonfiguracji reprezentacji i wyboru inwariantów

Rozwiązanie nie zawsze kryje się w dłuższym łańcuchu dedukcji; czasem pojawia się dzięki temu, że wcześniej błędnie określiliśmy, co właściwie powinno zostać poddane dedukcji. Z punktu widzenia filozofii nauki podobny mechanizm widać w zmianach teoretycznych. Nowa teoria nie musi jedynie dostarczać innych odpowiedzi na te same pytania – może również przekształcić katalog pytań, obiektów i relacji uznawanych za podstawowe. Przejście od fizyki arystotelesowskiej do mechaniki nowożytnej, od substancjalnych koncepcji ciepła do termodynamiki, od klasycznego pojęcia gatunku do populacyjnego myślenia biologii ewolucyjnej czy od prostych modeli racjonalnego podmiotu do koncepcji ograniczonej racjonalności nie polegało wyłącznie na dołożeniu nowych faktów. Zmieniała się ontologia robocza problemu. W skali mikro FARG próbuje modelować właśnie taką zdolność: system ma nie tylko sprawnie poruszać się wewnątrz swojego świata pojęciowego, ale – w ograniczonym zakresie – rekonfigurować jego lokalne granice. Jednocześnie teoria płynności nie może zostać pomyłona z relatywizmem. Gdyby każde pojęcie mogło dowolnie przechodzić w każde inne, analogia straciłaby wartość poznawczą.

Dobry system potrzebuje zarówno podatności na przesunięcie, jak i oporu wobec przesunięcia. Autor wyraża tę zasadę pojęciem głębokości conceptualnej: pewne relacje mają być bardziej stabilne niż inne, ponieważ reprezentują bardziej abstrakcyjne lub strukturalnie istotne własności problemu. W trafnej analogii przesuwają się zwykle elementy powierzchniowe, podczas gdy głębsze relacje zostają zachowane. Dzięki temu analogia nie jest swobodnym skojarzeniem, lecz operacją zachowującą wybrany inwariant przy zmianie pozostałych elementów. To właśnie pojęcie inwariantu pozwala połączyć FARG z szerszą teorią abstrakcji. Abstrakcja nie polega jedynie na usuwaniu szczegółów; jest procesem rozpoznawania tego, co powinno pozostać niezmiennie mimo zmiany konkretnej realizacji. Litera może mieć tysiące kształtów, a jednak być rozpoznawana jako ta sama litera; dwie historie mogą mieć zupełnie różnych bohaterów, a jednak realizować ten sam schemat relacyjny; rozwiązanie matematyczne może pojawiać się w pozornie odległych domenach, ponieważ zachowana zostaje struktura transformacji. Inteligencja abstrakcyjna wymaga zatem jednoczesnego rozpoznawania różnic i ignorowania ich części. Problem nie polega na maksymalnym podobieństwie, lecz na wyborze tego wymiaru podobieństwa, który jest istotny dla aktualnego zadania.

W tym miejscu koncepcja Hofstadtera zaczyna wykraczać poza historyczny spór z symboliczną AI. Jeżeli inteligencja zależy od zdolności wybierania odpowiednich inwariantów, reorganizowania reprezentacji i sterowania własnym poziomem eksploracji, to zasadnicze pytanie wobec każdego

systemu poznawczego brzmi nie tylko: „ile wie?”, ale również: „jak zmienia sposób organizowania tego, co wie?”. System może dysponować ogromnym zasobem informacji, a jednak być kruchy wobec niewielkiej zmiany formy zadania. Może też posiadać niewielką wiedzę, lecz imponującą zdolność rekonfiguracji reprezentacji. Wiedza i inteligencja nie są więc pojęciami tożsamymi, podobnie jak pamięć nie jest tożsama z rozumieniem, a poprawna odpowiedź – z kompetencją pozwalającą ją wytworzyć w nowych warunkach.

Na tym tle szczególnego znaczenia nabiera metoda mikrodomen. Z dzisiejszej perspektywy łatwo uznać problemy z literami, anagramami, stylizowanymi alfabetami czy prostymi ciągami liczbowymi za zbyt małe, by mogły powiedzieć coś istotnego o inteligencji. Takie zastrzeżenie jest częściowo zasadne: zachowanie w zamkniętej mikrodomenie nie gwarantuje skalowalności do otwartego świata, języka naturalnego, relacji społecznych czy działania ucieleśnionego. Równocześnie jednak mikrodomena daje korzyść metodologiczną. Pozwala obserwować niemal każdy element procesu, rozdzielić mechanizmy i sprawdzić, dlaczego określona interpretacja wygrała z konkurencyjną. W systemie obejmującym miliardy parametrów i ogromny korpus danych łatwiej uzyskać imponujący rezultat, ale znacznie trudniej odpowiedzieć na pytanie, jakie dokładnie procesy doprowadziły do jego powstania.

Paradoksalnie więc prostota modeli FARG może być ich ograniczeniem technologicznym, a jednocześnie przewagą epistemologiczną. Tę różnicę trzeba zachować w dalszej analizie współczesnych modeli językowych. Nie będziemy zakładać, że skoro LLM powstały z zupełnie odmiennej tradycji architektonicznej, Hofstadter został obalony. Nie będziemy też twierdzić odwrotnie, że ponieważ modele potrafią tworzyć analogie, muszą wewnątrz realizować mechanizmy podobne do Slipnetu i Coderacka. Oba wnioski byłyby zbyt pochopne.

Bardziej produktywne pytanie brzmi: jakie problemy funkcjonalne, które Hofstadter uważał za centralne dla inteligencji, zostały faktycznie rozwiązane przez współczesne systemy, w jaki sposób to zrobiono oraz które problemy pozostają nierozwiązane mimo imponującego zachowania? W ten sposób FARG przestaje pełnić rolę wzorca architektonicznego, który technologia powinna naśladować, a staje się aparatem pytań diagnostycznych.

Poszukiwanie wzorców jako proces zmiany reprezentacji

Pierwszym z tych pytań jest poszukiwanie wzorców. Dla Hofstadtera nie stanowi ono jednej z wielu specjalistycznych zdolności umysłu, lecz proces leżący u podstaw tworzenia pojęć, analogii i odkryć.

Umysł nie otrzymuje świata jako katalogu gotowych regularności; musi najpierw rozstrzygnąć, gdzie kończy się przypadkowe podobieństwo, a zaczyna struktura zasługująca na nazwę. Dlatego analiza musi zejść poniżej poziomu ogólnej filozofii płynności i przyjrzeć się samej anatomii odkrycia: przejściu od surowej sekwencji elementów, przez lokalne "wyspy porządku" i tymczasowe pakiety, ku takiej zmianie reprezentacji, dzięki której pozornie chaotyczny układ ujawnia własną zasadę organizacji. Dopiero w tym miejscu widać, dlaczego dla FARG inteligencja zaczyna się nie od odpowiedzi, lecz od umiejętności dostrzeżenia, na jakie pytanie naprawdę należy odpowiedzieć.

Płynność pojęciowa ma być czymś więcej niż metaforą elastyczności – musi zostać pokazana w działaniu. W programie badawczym FARG właśnie temu służyły eksperymenty z rozpoznawaniem struktur w ciągach liczbowych. Program Seek-Whence nie był projektem mającym jedynie przewidzieć kolejny element sekwencji. Jego znaczenie polegało na próbie modelowania procesu, dzięki któremu system stopniowo przestaje widzieć ciąg jako serię oddzielnych elementów, a zaczyna dostrzegać w nim organizację wyższego rzędu. Hofstadter opisuje ten proces jako oscylację między obliczaniem, zgadywaniem a rewizją hipotez, przy czym zasadniczym obiektem zmiany nie jest wyłącznie odpowiedź, lecz sposób segmentowania i interpretowania danych.

Taka perspektywa odpowiada jednej z najbardziej trwałych intuicji psychologii poznawczej: rozwiązanie problemu nie zawsze jest funkcją liczby operacji wykonanych na tej samej reprezentacji. Czasem wymaga ono powstania reprezentacji nowej. Psychologia postaci już w pierwszej połowie XX wieku zwracała uwagę na zjawisko nagłego wglądu, w którym układ elementów pozostaje fizycznie identyczny, lecz zmienia się jego poznawcza organizacja. Problem nie zostaje rozwiązany przez dodanie nowego faktu, ale przez zmianę relacji między tym, co już było dostępne.

Późniejsze teorie reprezentacyjnej zmiany rozwijały tę intuicję, wskazując, że impas może wynikać z nieadekwatnego sposobu kodowania ograniczeń zadania. System utrzymuje wtedy lokalnie poprawną, lecz globalnie jałową organizację problemu. Przełom wymaga rozluźnienia wcześniejszych założeń, przesunięcia granic grup lub zmiany cechy uznawanej dotąd za istotną. Seek-Whence można odczytać właśnie jako obliczeniową próbę uchwycenia takiego procesu; program nie powinien jedynie wykrywać częstotliwości, różnic arytmetycznych czy prostych okresowości.

Płynność poznawcza jako balans między dogmatyzmem a chaosem

System musi tworzyć tymczasowe struktury pośrednie — "wyspy porządku", pakiety i szablony — które następnie mogą zostać wzmacniane, modyfikowane albo porzucane. Autor książki definiuje "wyspy porządku" jako lokalne regularności, pakiety jako grupy elementów traktowane percepcyjnie jako jednostki, a szablony jako hipotetyczne struktury wyższego poziomu organizujące większy fragment ciągu. Znaczenie tych pojęć wykracza daleko poza analizę sekwencji; pokazują one, że rozumowanie wymaga pośrednich obiektów poznawczych, które nie są ani danymi pierwotnymi, ani końcową teorią. Są to konstrukcje tymczasowe, wystarczająco stabilne, by wspierać dalsze wnioskowanie, lecz na tyle elastyczne, aby można je było porzucić bez załamania całego systemu. Proces odkrycia ma zatem strukturę iteracyjną: najpierw pojawia się słabe przypuszczenie, że pewne elementy warto ze sobą związać, a następnie system sprawdza, czy lokalne grupowanie pomaga przewidywać kolejne elementy lub redukuje opisową złożoność danych.

Jeśli tak, rośnie poznawcza "inwestycja" w daną strukturę. W przypadku pojawienia się anomalii możliwe są co najmniej trzy reakcje: uznanie ich za przypadkowy szum, lokalna korekta hipotezy albo całkowita przebudowa reprezentacji. Inteligencja ujawnia się szczególnie w zdolności rozstrzygnięcia, kiedy wystarczy pierwsza lub druga reakcja, a kiedy niezbędna jest trzecia.

System zbyt sztywny będzie chronił dotychczasowy schemat przed falsyfikacją; system zbyt chwiejny porzuci każdą hipotezę po pierwszym odstępstwie. Płynność poznawcza jest więc umiejętnością utrzymania reprezentacji w stanie pośrednim między dogmatyzmem i chaosem. Zagadnienie to można opisać również w języku teorii uczenia i statystyki, gdzie każdy system poszukujący wzorców mierzy się z napięciem między dopasowaniem a generalizacją. Wzorzec może zostać dopasowany do niemal dowolnego skończonego zbioru danych, jeśli pozwolimy modelowi na wystarczającą złożoność, zatem sama zgodność z obserwacjami nie wystarcza, by uznać strukturę za interesującą. Potrzebne są dodatkowe preferencje: prostota, kompresyjność, regularność, możliwość ekstrapolacji, stabilność wobec drobnych zmian oraz zdolność ujawniania relacji niewidocznych w danych pierwotnych. W programie FARG preferencje te przyjmują postać swoistej estetyki wzorca — skłonności do uznawania za wartościowe struktur eleganckich, regularnych, rekurencyjnych i nieskażonych arbitralnymi wyjątkami. Autor książki wiąże tę funkcję z figurą Mathgoda i traktuje estetykę nie jako ornament, lecz jako mechanizm selekcji między konkurencyjnymi opisami.

Ten aspekt jest szczególnie interesujący z epistemologicznego punktu widzenia. Nauka również nie wybiera teorii wyłącznie na podstawie zdolności dopasowania do znanych danych; w praktyce znaczenie mają prostota, zakres wyjaśniania, płodność heurystyczna, unifikacja oraz odporność na mnożenie założeń pomocniczych. Nie oznacza to jednak, że kryterium piękna jest tożsame z kryterium prawdy — historia nauki dostarcza wielu przykładów eleganckich teorii, które okazały się błędne. Oznacza to natomiast, że proces poznawczy musi dysponować zasadami selekcji hipotez wykraczającymi poza mechaniczną zgodność z dotychczasową próbką. Hofstadterowska estetyka wzorca jest więc próbą modelowania intuicji, że nie wszystkie poprawne opisy są poznawczo równoważne.

Kluczowym przykładem w książce jest sekwencja określana jako "trójkąty między kwadratami". Punktem wyjścia są relacje między liczbami trójkątnymi i kwadratowymi, z których uzyskuje się ciąg dwójek i jedynek. Początkowo można próbować traktować go jako układ okresowy albo szukać niewielkich bloków powtarzających się lokalnie. Takie podejście przez pewien czas daje poczucie porządku, lecz kolejne elementy ujawniają jego niewystarczalność. Autor opisuje przejście od prostego odczytywania ciągu poprzez pakietyzację i rekonfigurację ku dostrzeżeniu rekurencyjnej struktury, w której operacja wykonywana na sekwencji generuje strukturę pozostającą w szczególnej relacji do sekwencji pierwotnej. Najważniejszy nie jest tu sam wzór matematyczny, lecz rodzaj transformacji poznawczej.

Na początku jedynki i dwójki są obiektami pierwszego rzędu. Następnie pewne ich układy stają się jednostkami drugiego rzędu, a potem relacje między tymi jednostkami zaczynają być traktowane jako nowe dane. System przestaje więc wyłącznie patrzeć na elementy ciągu, a zaczyna obserwować operację własnego grupowania. W ten sposób dochodzi do poziomu meta: relacja, która wcześniej służyła analizie danych, sama staje się przedmiotem analizy. Autor określa podobną operację mianem "drugiej pochodnej" i wiąże ją później z Metacatem oraz analogiami meta-poziomu.

Metapoznanie jako proces rekonfiguracji reprezentacji

W kategoriach kognitywnych mamy tu do czynienia z przejściem od reprezentowania świata do reprezentowania własnego sposobu reprezentowania świata. Nie musi ono oznaczać samoświadomości w sensie fenomenalnym, jest jednak początkiem metapoznania funkcjonalnego: system może traktować własne działania poznawcze jako dane dla kolejnych operacji. W ludzkim rozumowaniu podobna zdolność stanowi podstawę refleksji metodologicznej. Badacz nie tylko formułuje teorię, lecz potrafi analizować, dlaczego to zrobił, jakie heurystyki doprowadziły go do konkretnego modelu, gdzie pojawiły się impasy i które przesunięcia pojęciowe umożliwiły postęp.

Taki poziom meta pozwala na transfer strategii między domenami, ponieważ zapamiętywane jest nie tylko gotowe rozwiązanie, lecz również struktura drogi do rozwiązania. Seek-Whence okazuje się zatem istotny nie dlatego, że rozwiązuje konkretny ciąg liczbowy, lecz dlatego, że pozwala sformułować ogólną teorię poznawczego przejścia od danych do struktury.

Proces ten można rozłożyć na kilka funkcjonalnie odrębnych, choć wzajemnie sprzężonych etapów. System musi najpierw wykrywać lokalne regularności, a następnie tworzyć jednostki wyższego poziomu, które pozwalają zredukować liczbę elementów wymagających równoczesnego śledzenia. Musi oceniać konkurencyjne sposoby grupowania, rozpoznawać niewystarczalność dotychczasowego schematu, utrzymywać anomalie w polu uwagi i – gdy lokalna korekta zawodzi – dopuścić zmianę samego sposobu kodowania problemu. Dopiero wówczas możliwy staje się wgląd: nagłe powstanie reprezentacji, w której wcześniej rozproszone elementy stają się konsekwencją jednej reguły. Psychologia wglądu wskazuje, że proces ten jest często retrospektywnie zafałszowywany przez pamięć. Gdy rozwiązanie zostanie odnalezione, wcześniejsza trudność wydaje się niezrozumiała; nowa reprezentacja staje się tak oczywista, że trudno odtworzyć, dlaczego nie dostrzegaliśmy jej wcześniej. Hofstadter szczególnie interesował się właśnie tym bałaganem poprzedzającym elegancję rozwiązania. Odkrycie nie jest prostym marszem od przesłanek do wniosku; składa się z prób, błędów, niepełnych hipotez, lokalnych sukcesów, jałowych powtórzeń i reinterpretacji. Dopiero post factum umysł tworzy opowieść bardziej uporządkowaną niż rzeczywisty proces poznawczy. Ta intuicja powraca później w koncepcji temporalnego śladu Metacata, który ma przechowywać nie tylko wynik, lecz zgrubną historię drogi prowadzącej do wyniku.

Tak rozumiany wgląd podważa popularny model inteligencji jako czystej szybkości. System może błyskawicznie wykonywać operacje w ramach błędnej reprezentacji i pozostawać poznawczo bezradny. Inny system może długo oscylować między hipotezami, lecz dzięki zdolności przebudowy reprezentacji ostatecznie odkryć strukturę o większej mocy wyjaśniającej. Z tego punktu widzenia czas reakcji nie jest prostym miernikiem inteligencji – może on oznaczać nieefektywność, ale może również odzwierciedlać rzeczywistą pracę rekonfiguracji pojęciowej. To wyjaśnia, dlaczego Hofstadter przywiązywał wagę nie tylko do odpowiedzi, lecz również do „zmagania się” systemu z problemem i do rozkładu trudności ujawniającego się w procesie dochodzenia do rozwiązania. W tym miejscu pojawia się pierwsza poważna różnica między badaniem człowieka a oceną sztucznej inteligencji. W eksperymencie psychologicznym można obserwować czas rozwiązania, błędy, ruchy oczu, zmiany strategii, raporty introspekcyjne i transfer między zadaniami.

W modelu obliczeniowym można dodatkowo analizować wewnętrzny stan architektury, o ile została ona zaprojektowana w sposób interpretowalny. FARG cenił właśnie taką przejrzystość:

mikrodomena była wystarczająco mała, aby można było śledzić zmiany aktywacji, konkurencję struktur i moment powstawania rozwiązania.

Dynamiczna ziarnistość reprezentacji jako miara inteligencji

We współczesnych dużych modelach neuronowych sytuacja jest niemal odwrotna. Zakres zachowań jest nieporównanie większy, jednak pełne przyczynowe wyjaśnienie konkretnej odpowiedzi staje się znacznie trudniejsze. Otrzymujemy zatem systemy bardziej kompetentne behawioralnie, ale epistemicznie mniej przezroczyste. Różnica ta będzie miała zasadnicze znaczenie dla późniejszej analizy modeli językowych. Jeśli LLM poprawnie rozwiązuje analogię lub odkrywa wzorzec, możemy zmierzyć rezultat; znacznie trudniej jest jednak rozstrzygnąć, czy system dokonał czegoś funkcjonalnie odpowiadającego reprezentacyjnej zmianie, czy też wygenerował wynik dzięki innej strukturze przetwarzania.

W tym miejscu należy szczególnie wystrzegać się antropomorfizmu. Fakt, że odpowiedź przypomina rezultat ludzkiego wglądu, nie dowodzi tożsamości procesu. Analogicznie jednak odmienność architektury nie świadczy o braku funkcjonalnie równoważnego mechanizmu. Nauka potrzebuje tutaj testów pośrednich: badań nad transferem, odpornością na parafrazę, zachowaniem po zmianie formatu, generalizacją poza rozkład treningowy, reakcją na kontrfaktyczności oraz analizy zmian wewnętrznych reprezentacji.

Seek-Whence można potraktować jako wcześniejszą wersję problemu generalizacji strukturalnej. System nie powinien jedynie zapamiętywać konkretnych sekwencji, lecz uchwycić zasadę umożliwiającą rozpoznanie podobnej organizacji w nowych przypadkach. Jest to różnica między rozpoznaniem egzemplarza a rozpoznaniem relacji. Człowiek dostrzega analogię między układem liczb, rytmem muzycznym, strukturą narracji czy problemem przestrzennym nie dlatego, że elementy tych domen są powierzchownie podobne, lecz ponieważ można odwzorować między nimi stosunki wyższego rzędu. Im wyższy poziom abstrakcji, tym większe znaczenie ma zdolność odrywania relacji od ich materialnej realizacji. Współczesna teoria analogii często odróżnia podobieństwo atrybutowe od relacyjnego. Dwa obiekty mogą wyglądać podobnie, lecz pełnić odmienne role; dwa zupełnie różne obiekty mogą natomiast pozostawać w analogicznych relacjach wobec innych elementów struktury. Inteligentna analogia wymaga nie tylko znalezienia odpowiadających sobie elementów, ale ustalenia, które relacje należy zachować. Problem ten stanie się później centralny w Copycat.

Tam system zostanie zmuszony rozstrzygnąć, czy ważniejsza jest literalna tożsamość elementów, ich położenie w alfabecie, kierunek relacji, czy funkcja strukturalna. Seek-Whence przygotowuje grunt pod tę teorię, pokazując, że elementy zyskują poznawcze znaczenie dopiero dzięki sposobowi, w jaki są grupowane. Interesujący jest również związek tej koncepcji z ograniczeniami pamięci roboczej. Człowiek nie może jednocześnie śledzić arbitralnie dużej liczby niezależnych elementów, dlatego jednym z podstawowych mechanizmów radzenia sobie z tym limitem jest chunking, czyli grupowanie danych w jednostki wyższego rzędu. Znany numer telefonu nie jest przechowywany jako lista niepowiązanych cyfr, podobnie jak doświadczony szachista nie postrzega pozycji wyłącznie jako zbioru oddzielnych figur.

Wiedza zmienia sposób segmentowania materiału. Seek-Whence wykorzystuje podobną ideę, nadając jej jednak charakter dynamiczny: pakiety nie są z góry przygotowanymi jednostkami, lecz kandydatami na jednostki, które mogą zostać przebudowane pod wpływem nowych danych. Statyczny chunking kompresuje informacje zgodnie z wcześniej poznany schematem, podczas gdy płynna pakietyzacja uczestniczy w odkrywaniu samego schematu. System nie wie jeszcze, co jest właściwą jednostką analizy; dopiero przebieg problemu ujawnia, czy bardziej produktywnie jest traktowanie sekwencji `2`, `1`, `2` jako trzech osobnych obiektów, elementu większego pakietu, markerów granic czy wyników operacji na strukturze wyższego poziomu.

Inteligencja nie ogranicza się więc do efektywnego przechowywania informacji, lecz obejmuje zdolność wyboru odpowiedniej „ziarnistości” reprezentacji. Pojęcie to prowadzi do kolejnej ważnej własności poznania: ten sam obiekt może być sensownie reprezentowany na wielu poziomach jednocześnie. Litera jest układem linii, znakiem alfabetu, elementem słowa, szyfru albo punktem w przestrzeni relacji porządkowych. Liczba może być wielkością, pozycją w sekwencji, wynikiem działania, elementem symetrii lub reprezentantem klasy. Nie istnieje zatem jeden neutralny poziom opisu właściwy dla wszystkich zadań; kompetentny system musi przełączać się między nimi zależnie od aktualnego celu. Ta zdolność do zmiany rozdzielczości pojęciowej jest jednym z najważniejszych aspektów płynności. W tym miejscu program FARG można zestawić z szerszymi koncepcjami predykcyjnego poznania, choć nie należy ich utożsamiać. W teoriach predictive processing percepcja jest często ujmowana jako proces wzajemnego oddziaływania przewidywań wyższego poziomu i sygnałów błędu z poziomów niższych. System utrzymuje model generatywny świata i modyfikuje go, gdy przewidywania systematycznie rozmiągają się z danymi. W FARG również występuje dynamiczne sprzężenie top-down i bottom-up, jednak formalna architektura oraz cele teoretyczne są odmienne.

Błąd jako sygnał do rekonfiguracji reprezentacji

Wspólną intuicją jest odrzucenie naiwnego modelu percepcji jako biernego odbioru danych. To, co zostaje „zobaczone”, zależy od oczekiwań, kontekstu, hipotez i lokalnych niezgodności. Anomalia ma w takim systemie szczególną wartość – nie jest wyłącznie błędem do usunięcia, lecz sygnałem, że aktualna reprezentacja stała się niewystarczająca. W historii nauki podobną funkcję pełniły obserwacje traktowane początkowo jako problemy lokalne, aż ich nagromadzenie wymusiło zmianę szerszej struktury teoretycznej. Nie należy jednak mechanicznie przenosić Kuhnowskiego pojęcia rewolucji naukowej na mikrodomeny FARG.

Analogia ta działa na innym poziomie, lecz w obu przypadkach kluczowe jest to, że dane nie wymuszają automatycznie jednej interpretacji; ich znaczenie zależy od sieci pojęć, w której są osadzone. Tutaj ujawnia się słabość czysto indukcyjnego obrazu odkrycia. Gdyby poznanie sprowadzało się do prostego gromadzenia przykładów i liczenia częstotliwości, wystarczyłaby ekspozycja na większą liczbę danych. Tymczasem nowe informacje często nie pomagają, dopóki nie zmieni się sposób ich organizacji. Można wielokrotnie patrzeć na te same elementy, nie dostrzegając struktury, która po odpowiedniej reinterpretacji staje się nagle oczywista. To dlatego wgląd bywa subiektywnie odczuwany jako skok. Nie musi on jednak oznaczać procesu magicznego czy niezdeterminowanego. Nagłość fenomenologiczna może być końcowym skutkiem stopniowej akumulacji lokalnych zmian, które po przekroczeniu pewnego progu stabilizują nową reprezentację.

Takie ujęcie zbliża wgląd do zjawisk nieliniowych. Przez pewien czas wiele słabych presji może nie zmieniać dominującej interpretacji, jednak kolejne anomalie, aktywacje i alternatywne grupowania stopniowo osłabiają istniejącą strukturę. Gdy przewaga konkurencyjnej organizacji staje się wystarczająca, system przechodzi do nowego stanu, co z perspektywy obserwatora wygląda jak nagły przełom.

Jest to istotne również dla rozumienia twórczości. Kreatywność może być fenomenologicznie skokowa, choć mechanizm prowadzący do niej pozostaje rozproszony i stopniowy. FARG wykorzystuje tę intuicję konsekwentnie: geniusz i błąd nie muszą wynikać z dwóch różnych mechanizmów. Ta sama płynność, która pozwala znaleźć nieoczywistą analogię, może doprowadzić do błędnego skojarzenia. Autor przywołuje tu przykład omyłkowej frazy „once in a bloom”, powstałej przez nakładanie się wpływów fonetycznych i kontekstowych.

Sens tej obserwacji wykracza poza anegdotę językową: system, który nigdy nie popełnia błędów wynikających z interferencji i reinterpretacji, może być jednocześnie zbyt sztywny, by osiągać kreatywne przesunięcia. Nie oznacza to oczywiście, że większa liczba pomyłek świadczy o wyższej

inteligencji – kluczowa jest struktura błędu. Błąd losowy nie mówi wiele o organizacji poznawczej; błąd systematyczny może natomiast ujawnić, jakie relacje były aktywne, które podobieństwa uznano za istotne i jak konkurencyjne reprezentacje oddziaływały na siebie. W psychologii poznawczej analiza pomyłek od dawna służy jako źródło informacji o mechanizmach przetwarzania. Przejęzyczenia, błędy pamięciowe, złudzenia percepcyjne i błędne analogie odsłaniają architekturę systemu właśnie dlatego, że pokazują, w jakich warunkach zawodzą jego standardowe heurystyki. Z tej perspektywy prawdziwym przeciwieństwem inteligencji nie jest błąd, lecz niezdolność do wykorzystania go jako informacji. System inteligentny może się pomylić, ale powinien posiadać mechanizmy wykrywania niespójności, rewizji i uczenia się na podstawie własnej historii.

Copycat jako laboratorium dynamicznej rekonfiguracji pojęć

W FARG ta intuicja prowadzi od Seek-Whence do Metacata. Zanim jednak pojawi się pełniejszy poziom metapoznawczy, konieczne jest zrozumienie mechanizmu umożliwiającego samą zmianę reprezentacji. Jest nim Copycat: architektura, w której pojęcia nie mają niezmiennych odległości, interpretacje konkurują ze sobą, a lokalne porażki mogą zwiększać prawdopodobieństwo radykalniejszych przesunięć pojęciowych.

W ten sposób Seek-Whence pełni w programie FARG rolę metodologicznego prologu. Pokazuje on, że nawet prosta sekwencja liczb staje się interesująca dopiero wtedy, gdy pytamy nie tylko o to, jaka jest reguła, lecz również o to, jak system dochodzi do tego, co w ogóle uzna za regułę. To drugie pytanie jest trudniejsze, gdyż dotyczy genezy reprezentacji, selekcji hipotez oraz granicy między przypadkiem a strukturą. Nie da się na nie odpowiedzieć poprzez proste dołożenie kolejnej tabeli faktów.

Wymaga to teorii poznania zdolnej opisać proces przechodzenia od elementów do relacji, od relacji do relacji między relacjami i od lokalnego impasu do reorganizacji całego pola problemowego. Dopiero na tym tle można właściwie zrozumieć znaczenie Copycata. Jego słynne analogie literowe nie są intelektualną zabawą w alfabet, lecz skrajnie uproszczonym laboratorium konfliktu między podobieństwem powierzchownym i głębokim, stabilnością i poślizgiem, lokalnym dopasowaniem i globalną koherencją. Jeżeli Seek-Whence dowodzi, że odkrycie wymaga reorganizacji sposobu grupowania danych, Copycat stawia pytanie jeszcze bardziej wymagające: co powinno pozostać niezmiennie, kiedy wszystko inne można zmienić? Odpowiedź na nie prowadzi bezpośrednio do teorii analogii, abstrakcji i elastycznej reprezentacji. Stamtąd już tylko krok do najtrudniejszego

problemu współczesnych modeli językowych: czy system potrafiący tworzyć przekonujące analogie w tysiącach domen rzeczywiście rozpoznaje relacje strukturalne, czy jedynie niezwykle sprawnie porusza się po statystycznym krajobrazie form. Projekt Copycat zajmuje w architekturze intelektualnej FARG miejsce szczególne, ponieważ jednoczy większość intuicji obecnych wcześniej w Seek-Whence, Jumbo i Numbo w jeden model wysokopoziomowej percepcji.

Jego domena jest pozornie niemal groteskowo prosta: krótkie ciągi liter alfabetu i transformacje w rodzaju $abc \rightarrow abd$, do których należy znaleźć analogiczne przekształcenie innego ciągu. Prostota ta jest jednak metodologicznie zamierzona. Problem ma być na tyle ograniczony, aby niemal każdą istotną zmianę reprezentacji można było zaobserwować, a zarazem wystarczająco bogaty, by ujawniły się konflikty między różnymi sposobami rozumienia podobieństwa. Copycat nie jest więc testem znajomości alfabetu; to eksperyment badający, jak system ustala, które własności sytuacji są przypadkowe, które istotne, a które można poświęcić w celu zachowania głębszej relacji.

Architektura Copycata jako system dynamicznej reprezentacji

Autor przedstawia Copycat jako architekturę zbudowaną wokół trzech współdziałających elementów: Slipnetu, Workspace oraz Coderacka. Slipnet jest siecią pojęć o strukturze pozbawionej całkowitej sztywności. Workspace stanowi przestrzeń roboczą, w której tymczasowo budowane są grupy, relacje i mosty pomiędzy elementami analizowanych ciągów. Coderack przechowuje natomiast populację kodletów – drobnych procedur wykonujących lokalne zadania eksploracyjne i konstrukcyjne. System nie posiada centralnego dyspozytora, który najpierw „rozumie” zadanie, a następnie nakazuje pozostałym modułom realizację gotowego planu.

Interpretacja wyłania się stopniowo z interakcji tych trzech warstw. Najbardziej charakterystycznym składnikiem jest Slipnet. Nazwa oddaje istotę modelu: to sieć, po której pojęcia mogą się „ślizgać”. Nie należy jednak rozumieć tego jako dowolnej zamienności znaczeń. Sieć reprezentuje relacje takie jak „następnik”, „poprzednik”, „pierwszy”, „ostatni”, „prawo”, „lewo”, „tożsamość”, „przeciwnieństwo” czy różne poziomy abstrakcji. Pod wpływem aktualnego kontekstu pewne pojęcia mogą zostać silniej aktywowane, inne słabiej, a odległości funkcjonalne między nimi ulegają zmianie.

Dzięki temu system nie jest skazany na jedną literalną interpretację zadania. Jeżeli pierwotne odwzorowanie prowadzi do impasu, kontekst może zwiększyć dostępność pojęć wcześniej peryferyjnych i otworzyć możliwość alternatywnej analogii. Autor opisuje to za pomocą dwóch

pojęć: **conceptual halos** (pojęciowych otoczek) oraz **conceptual slippage**, czyli kontrolowanego poślizgu pomiędzy kategoriami. W tym rozwiązywaniu kryje się ważna intuicja dotycząca semantyki. Znaczenie pojęcia nie jest wyłącznie jego definicją słownikową ani pozycją w niezmienniej taksonomii, lecz funkcją relacji, które stają się aktywne w danym kontekście. Pojęcie „ostatni” może być mało istotne podczas rozwiązywania jednego zadania, a centralne w innym. Podobnie „przeciwieństwo” może nie należeć do domyślnej interpretacji relacji między dwoma elementami, lecz zostać aktywowane wtedy, gdy literalne odwzorowanie okazuje się niemożliwe. Semantyka Copycata jest zatem częściowo relacyjna i dynamiczna: pojęcia posiadają względnie stabilną strukturę, ale ich chwilowa dostępność oraz funkcja zależą od całego pola problemowego. Nie oznacza to jednak, że Slipnet stanowi ogólną teorię ludzkiej pamięci semantycznej – jego domena jest zbyt wąska, a pojęcia zostały zaprojektowane przez człowieka.

Jest to istotne ograniczenie, do którego trzeba powrócić w dyskusji o przypisywaniu zasług. Niemniej jako model funkcjonalny Slipnet pozwala uchwycić problem kluczowy niezależnie od konkretnej architektury: jak system może korzystać z kategorii, nie traktując ich granic jako absolutnie nieprzekraczalnych? Każda teoria elastycznego rozumowania musi odpowiedzieć na to pytanie. System, który nigdy nie opuszcza literalnej kategorii, jest poznawczo kruchy; system, który zmienia kategorie bez ograniczeń, traci możliwość stabilnego wnioskowania. Inteligencja musi więc operować w przestrzeni pomiędzy stałością a plastycznością.

Drugim składnikiem Copycata jest Workspace – przestrzeń robocza, w której konstruowane są chwilowe interpretacje aktualnego zadania. Autor porównuje ją do cytoplazmatycznego placu budowy: elementy nie są tam jedynie biernie przechowywane, lecz zostają wiązane, grupowane, rozdzielane i łączone mostami analogicznymi. Metafora ta dobrze oddaje epistemologiczny charakter reprezentacji w FARG. Reprezentacja nie jest gotowym obrazem świata, który system otrzymuje przed rozpoczęciem rozumowania, lecz strukturą roboczą, stale podlegającą rekonstrukcji. Część jej elementów może zostać zburzona, jeśli nowe relacje okażą się silniejsze; inne zostają wzmocnione, gdy pasują do coraz bardziej spójnej interpretacji całego problemu.

To właśnie w Workspace rozgrywa się zasadniczy konflikt między lokalnym a globalnym dopasowaniem. Lokalna relacja może wyglądać atrakcyjnie, lecz po włączeniu jej do większej struktury ujawnia niespójność. Z drugiej strony globalnie elegancka interpretacja może wymagać zaakceptowania lokalnego odwzorowania, które na początku wydaje się mniej oczywiste. Inteligentne tworzenie analogii wymaga więc przechodzenia pomiędzy tymi poziomami. Nie można oceniać pojedynczej korespondencji niezależnie od całej struktury, ani narzucać struktury globalnej, ignorując sygnały oddolne. Copycat próbuje rozwiązać ten problem przez ciągłe sprzężenie między propozycjami lokalnych relacji a stanem powstającej interpretacji. Trzeci

element, Coderack, odpowiada za generowanie i selekcję mikrodziałañ. Kodlety mogą badać potencjalne relacje, próbować tworzyć grupy, budować mosty analogiczne albo rozbijać struktury, które okazują się słabe.

Równoległy skan tarasowy jako mechanizm zarządzania budżetem poznawczym

Ich uruchamianie ma charakter probabilistyczny, jednak prawdopodobieństwo nie jest tu kwestią czystego przypadku – zależy ono od bieżącej aktywacji, pilności oraz stanu Workspace. W efekcie system nie realizuje jednego, zaprogramowanego ciągu kroków; wiele kierunków pozostaje przez pewien czas możliwych, choć nie wszystkie otrzymują taką samą ilość zasobów.

Architekturę tę można interpretować jako próbę uniknięcia dwóch skrajności. Pierwszą byłoby wyczerpujące przeszukiwanie wszystkich dostępnych interpretacji, co szybko doprowadziłoby do eksplozji kombinatorycznej. Drugą – całkowicie deterministyczny algorytm, który zawsze wybiera jedną ścieżkę na podstawie lokalnej heurystyki. Autor książki proponuje rozwiązanie pośrednie: eksploracja jest wielotorowa, lecz nierównomierna. Obiecujące kierunki przyciągają więcej mikrodziałañ, jednak alternatywy nie znikają natychmiast. Dzięki temu system może przez pewien czas zachować pluralizm interpretacyjny i dopiero później ustabilizować dominującą strukturę.

Mechanizm ten określono mianem **parallel terraced scan**, czyli równoległego skanu tarasowego. Jego istotą jest nierównomierne pogłębianie wielu hipotez. Kierunek, który zaczyna wyglądać obiecująco, otrzymuje więcej zasobów, ale mniej prawdopodobne ścieżki nie są od razu eliminowane. Z perspektywy teorii racjonalności można stwierdzić, że system dynamicznie zarządza ograniczonym budżetem poznawczym. Nie dysponując możliwością pełnego przeliczenia przestrzeni stanów, inwestuje uwagę proporcjonalnie do dotychczasowych sygnałów jakości. Jest to obliczeniowy odpowiednik heurystycznego rozdzielania uwagi przez człowieka.

Kluczowym regulatorem tej dynamiki jest temperatura. Wysoka temperatura oznacza, że system nie posiada jeszcze stabilnej interpretacji; w takich warunkach większa losowość i tolerancja wobec odległych skojarzeń zwiększają szansę na znalezienie alternatywnej reprezentacji. Gdy temperatura spada – ponieważ Workspace staje się bardziej spójny – zachowanie systemu ulega stabilizacji. Wówczas rośnie prawdopodobieństwo działań zgodnych z dominującą strukturą, a eksploracja staje się bardziej konserwatywna. Pozwala to sformułować ogólniejszą zasadę poznawczą: stopień elastyczności reprezentacji powinien zależeć od jakości aktualnego rozwiązania. Gdy interpretacja działa dobrze, nie ma powodu, by nieustannie ją destabilizować. Jeśli jednak rośnie liczba

konfliktów i niespójności, koszt utrzymywania dotychczasowej struktury staje się zbyt wysoki, a system powinien zwiększyć gotowość do radykalnych przesunięć. Analogiczny mechanizm obserwujemy w ludzkim rozwiązywaniu problemów: utrzymujemy robocze założenia tak długo, jak pozostają użyteczne, lecz trwały impas może skłonić nas do zakwestionowania nawet tych rozróżnień, które wcześniej wydawały się oczywiste.

Poślizg pojęciowy jako wybór inwariantu

Najbardziej znanym przykładem działania Copycata jest analogia $\text{'abc} \rightarrow \text{abd}; \text{xyz} \rightarrow \text{'}$. Na poziomie powierzchniowym transformacja pierwszego ciągu wydaje się banalna: ostatnia litera zostaje zastąpiona przez swojego następnika. Literalne zastosowanie tej reguły do 'xyz prowadzi jednak do impasu – w standardowym alfabecie po literze 'z nie ma już żadnej innej. W systemie sztywnym oznaczałoby to zwykłą niemożność wykonania operacji lub konieczność wprowadzenia specjalnego wyjątku. W Copycat ten opór ma pełnić inną funkcję: zwiększa presję na reorganizację sposobu interpretowania problemu.

Hofstadter opisuje ten proces sekwencyjnie. Najpierw pojawia się blokada, ponieważ standardowa relacja „następnik” nie znajduje zastosowania. Następnie wzrasta aktywacja otoczki pojęciowej związanej z literą 'z , dzięki czemu znaczenia takie jak „ostatnia w alfabecie” stają się bardziej dostępne. Pozwala to dostrzec strukturalną symetrię między 'a jako początkiem alfabetu a 'z jako jego końcem.

W konsekwencji możliwy staje się poślizg pojęciowy: „następnik” zostaje zastąpiony przez „poprzednika”, a kierunek transformacji ulega odwróceniu. Rozwiązaniem jest wówczas $\text{'xyz} \rightarrow \text{wyz}$. Nie sama odpowiedź jest tu jednak najciekawsza, gdyż zależnie od przyjętych kryteriów można by zaproponować wiele innych wariantów. Kluczowe jest to, że analogia nie zostaje potraktowana jako literalne kopiowanie operacji. System dąży do zachowania bardziej abstrakcyjnej relacji: zmiany elementu znajdującego się na szczególnej pozycji względem granicy alfabetu. Aby to osiągnąć, musi poświęcić literalną tożsamość operacji „następnik” na rzecz wyższego poziomu symetrii. To właśnie jest sednem koncepcji poślizgu: dobra analogia zachowuje nie wszystko, lecz właściwe rzeczy. Rozwiązywanie analogii wymaga zatem wyboru inwariantu.

Gdy porównujemy dwa układy, zawsze występują między nimi zarówno podobieństwa, jak i różnice. Gdyby wymagane było zachowanie wszystkich własności, prawdziwa analogia byłaby możliwa tylko między identycznymi obiektami. Z kolei gdyby można było zignorować dowolne cechy, każdy obiekt można by uznać za analogiczny do każdego innego. Poznawcza wartość analogii polega więc na selekcji relacji, które powinny zostać zachowane, oraz tych, które można

przekształcić. Copycat modeluje ten proces jako wynik konkurencji różnych nacisków pojęciowych, a nie jako zastosowanie jednej uniwersalnej definicji podobieństwa.

Mechanizm ten ma bezpośredni odpowiednik w psychologicznych teoriach analogii, które rozróżniają podobieństwo powierzchniowe od strukturalnego. Historie powierzchniowo podobne mogą opisywać te same przedmioty lub posługiwać się zbliżonym słownictwem, lecz realizować zupełnie inne relacje przyczynowe. Z kolei przypadki powierzchniowo odległe mogą posiadać identyczną strukturę relacyjną. Klasyczne eksperymenty nad transferem analogicznym pokazują, że ludzie często mają trudność z wykorzystaniem rozwiązania poznanego w jednej domenie, jeśli nowy problem różni się od niego powierzchniowo – nawet gdy relacje strukturalne pozostają niemal identyczne.

Percepcja jako proces aktywnego konstruowania reprezentacji

Samo posiadanie właściwej zasady nie gwarantuje transferu; trzeba jeszcze rozpoznać, że aktualna sytuacja należy do tej samej klasy relacyjnej. Właśnie tutaj spotykają się analogia i percepcja wysokiego poziomu. Aby odwzorować dwie struktury, system musi najpierw zdecydować, jakie właściwie struktury w nich istnieją. Nie można oddzielić mapowania od reprezentacji, ponieważ wybór tej drugiej decyduje o tym, które odwzorowania w ogóle stają się dostępne. Autor książki wyraźnie przeciwstawia w tym punkcie Copycata modelom, w których reprezentacja zostaje przygotowana przed rozpoczęciem analogii. W podejściu FARG konstruowanie reprezentacji oraz znajdowanie analogii są jednym sprzężonym procesem.

Teza ta jest znacznie bardziej radykalna, niż mogłoby się początkowo wydawać. Jeśli reprezentacja zależy od poszukiwanej analogii, to percepcja nie jest neutralną operacją poprzedzającą myślenie. Sposób zobaczenia sytuacji częściowo zależy od tego, do czego próbujemy ją porównać. Człowiek patrzący na przedmiot może widzieć go jako narzędzie, przeszkodę, znak, część większego układu albo egzemplarz kategorii – zależnie od aktualnego celu. Nie oznacza to, że rzeczywistość staje się dowolna; oznacza natomiast, że poznawcza artykulacja rzeczywistości jest zadaniowo zależna. Z tego punktu widzenia Copycat zbliża się do pojęcia **seeing as** – „widzenia jako”. Ten sam element fizyczny lub symboliczny może zostać rozpoznany pod różnymi opisami, z których każdy uwydatnia inne relacje.

W analizie analogii rozróżnienie między „widzeniem `z` jako po prostu litery `z`” a „widzeniem `z` jako końca uporządkowanego systemu” może przesądzić o możliwości rozwiązania problemu.

Inteligencja polega więc nie tylko na wiedzy o cechach elementu, lecz na zdolności aktywowania tego opisu, który jest właściwy w danej strukturze. Tak rozumiana elastyczność ma ścisły związek z abstrakcją. Abstrakcja nie jest jedynie redukcją szczegółów, lecz selekcją poziomu, na którym dwie sytuacje stają się porównywalne. Pozostając zbyt blisko detali, dostrzeżemy wyłącznie różnice. Odejście zbyt daleko sprawi natomiast, że analogia stanie się pusta, gdyż niemal wszystko będzie podobne do wszystkiego pod wystarczająco ogólnym opisem. Trafna abstrakcja wymaga odnalezienia poziomu pośredniego: na tyle ogólnego, by przekroczyć różnice powierzchniowe, lecz na tyle konkretnego, by zachować informację strukturalną.

W tym miejscu łatwiej zrozumieć Hofstadterowską zasadę, według której „głębokie” pojęcia powinny ślizgać się mniej niż powierzchowne. Autor książki ujmuje głębokość conceptualną jako stopień stabilności pojęcia podczas budowania dobrej analogii. Nie chodzi tu o absolutną hierarchię ontologiczną, w której pewne pojęcia są zawsze głębsze od innych, lecz o funkcjonalną zasadę zachowywania inwariantów: gdyby wszystkie elementy struktury były równie łatwo wymienialne, analogia utraciłaby treść. Kreatywny poślizg wymaga zatem stabilnego punktu odniesienia.

Można tu dostrzec podobieństwo do matematycznego pojęcia transformacji zachowującej określone własności. Geometria bada między innymi to, które cechy pozostają niezmiennie przy konkretnych klasach przekształceń. Analogia poznawcza nie jest oczywiście formalną izometrią ani homomorfizmem w ścisłym sensie matematycznym, lecz idea inwariancji pomaga zrozumieć jej logikę. Dobra analogia może silnie przekształcać elementy powierzchniowe, zachowując relację wyższego rzędu. Dzięki temu pozwala przenieść wiedzę między domenami, które na pierwszy rzut oka nie mają ze sobą wiele wspólnego. Analogia staje się więc mechanizmem generalizacji.

Gdy człowiek uczy się wyłącznie przez podobieństwo powierzchniowe, jego wiedza pozostaje związana z konkretnym kontekstem. Dopiero rozpoznanie struktury umożliwia transfer do nowej sytuacji. W edukacji ma to kluczowe znaczenie: student może znać dziesiątki rozwiązanych przykładów i nadal być bezradny wobec problemu sformułowanego inaczej, jeżeli nie wyodrębnił zasady organizującej. Ekspert różni się od nowicjusza między innymi tym, że częściej kategoryzuje problemy według głębokich mechanizmów, podczas gdy ten drugi łatwiej ulega podobieństwom powierzchniowym.

Copycat próbuje modelować właśnie przejście od jednego rodzaju podobieństwa do drugiego. Jednocześnie architektura ta ujawnia własne ograniczenia. Zawartość Slipnetu nie powstaje z próżni – projektanci określili pojęcia, relacje i podstawową organizację mikrodomeny. System posiada płynność w obrębie przygotowanego środowiska conceptualnego, lecz nie tworzy własnej ontologii. To ograniczenie staje się szczególnie istotne przy próbie wykorzystania Copycata jako

modelu ogólnej inteligencji. Można bowiem zapytać, czy system naprawdę „odkrywa”, że `z` jest końcem alfabetu, skoro relacja ta istnieje w skonstruowanej przez człowieka sieci pojęciowej.

Odpowiedź wymaga ostrożności: Copycat modeluje autonomię procesu wyboru i rekonfiguracji spośród dostępnych pojęć, a nie pełną autonomię genezy tych pojęć. Nie unieważnia to projektu, lecz określa jego właściwy zakres. Eksperyment pokazuje, jak dynamiczna sieć konceptualna może uczestniczyć w elastycznym rozwiązywaniu analogii, ale nie wyjaśnia jeszcze, jak kompletny system poznawczy nabywa swój aparat pojęciowy bezpośrednio ze świata.

Ta luka stanie się niezwykle interesująca w zestawieniu ze współczesnym uczeniem reprezentacji. Dzisiejsze modele neuronowe nie otrzymują od programisty jawnego Slipnetu zawierającego pojęcia „następnik”, „przeciwieństwo” czy „ostatni”. Ich wewnętrzne reprezentacje powstają podczas optymalizacji na ogromnych zbiorach danych. Nie oznacza to jednak automatycznie, że nabywają one pojęcia w takim sensie, w jakim rozumie je psychologia czy filozofia umysłu. Obie tradycje rozwiązują problem genezy reprezentacji w odmienny sposób i żadnej nie wolno zbyt szybko tłumaczyć na język drugiej. Szczególnie kuszące byłoby utożsamienie Slipnetu ze współczesną przestrzenią embeddingów.

Funkcjonalna analogia zamiast tożsamości **architektonicznej**

W obu przypadkach mamy do czynienia z reprezentacjami relacyjnymi, podobieństwami i zależnościami od sąsiedztwa. Choć analogia ta może być heurystycznie użyteczna, nie oznacza ona tożsamości mechanizmów. Slipnet został zaprojektowany jako jawna, interpretowalna sieć konceptualna, w której węzły pełnią określone role semantyczne, a ich aktywacja steruje konkretnymi procesami. Embedding w sieci neuronowej jest natomiast rozproszoną reprezentacją numeryczną, wyuczoną w procesie optymalizacji; pojedynczy wymiar rzadko odpowiada jednej jawnej kategorii, a geometria przestrzeni może kodować wiele nakładających się regularności. Podobieństwo dotyczy zatem funkcjonalnego problemu relacyjności reprezentacji, a nie technicznej identyczności rozwiązania.

Podobną ostrożność należy zachować przy porównywaniu Coderacka z mechanizmami współczesnych modeli. Fakt, że zarówno w FARG, jak i w dużych sieciach występuje rozproszenie przetwarzania, nie świadczy o wspólnej architekturze poznawczej. Kodlety są funkcjonalnie wyspecjalizowanymi mikroagentami w jawnie zaprojektowanym systemie, podczas gdy operacje transformera wynikają z wielowarstwowych przekształceń reprezentacji numerycznych wyuczonych

gradientowo. Kluczowe pytanie nie powinno zatem brzmieć: „gdzie w LLM znajduje się Coderack?”, lecz: „czy istnieją w LLM mechanizmy funkcjonalnie umożliwiające konkurencję interpretacji, zmianę reprezentacji, selekcję poziomu abstrakcji i rekonfigurację po napotkaniu blokady?”.

Właśnie takie pytania pozwalają zachować wartość FARG bez przekształcania jej w historyczną ortodoksję. Teoria nie musi zostać potwierdzona przez późniejszą technologię w szczegółach architektonicznych, aby pozostać intelektualnie płodna. Może dostarczać problemów, kryteriów i rozróżnień, które okazują się użyteczne wobec systemów zbudowanych zupełnie inaczej. Najtrwalszym dziedzictwem Copycata nie musi być zatem Slipnet ani temperatura jako konkretne elementy programu. Może nim być teza, że analogii nie da się oddzielić od procesu konstruowania reprezentacji, a inteligencji nie da się wiarygodnie ocenić, jeżeli bada się wyłącznie końcowy rezultat odwzorowania. Stąd wynika również krytyka prostego behawioryzmu w ocenie AI.

Jeżeli dwa systemy podają tę samą odpowiedź, nie oznacza to jeszcze, że posiadają tę samą kompetencję. Jeden może stosować wyuczoną regułę pasującą do znanego formatu, drugi zaś konstruować relację zdolną do transferu między domenami. Aby rozróżnić te możliwości, należy zmieniać kontekst, format, powierzchnię struktury zadania i dostępne wskazówki. System rozumiejący relację powinien w większym stopniu zachowywać kompetencję mimo takich zmian; system oparty przede wszystkim na lokalnych korelacjach może natomiast wykazywać znacznie większą kruchość. Takie podejście kieruje analizę w stronę problemu kontrfaktyczności. Hofstadter interesował się zadaniami pozornie dziwacznymi właśnie dlatego, że zmuszają one system do opuszczenia najłatwiejszego toru skojarzeń. Pytanie „co byłoby, gdyby 2 było 3?” nie ma służyć sprawdzeniu wiedzy arytmetycznej.

Inteligencja jako dynamiczne zarządzanie reprezentacjami i niepewnością

Celem jest zbadanie, czy system potrafi tymczasowo zawiesić jedną ustaloną relację i przeorganizować sieć konsekwencji wokół alternatywnego założenia. W rzetelnym eksperymencie kontrfaktycznym nie chodzi o to, czy odpowiedź zgadza się z rzeczywistością, lecz czy system zachowuje spójność w nowym, świadomie zmodyfikowanym świecie pojęciowym. Tak rozumiane kontrfaktyczności stanowią szczególnie wartościowy test abstrakcji. System opierający się jedynie na statystycznych regularnościach konkretnego środowiska łatwo traci spójność po zmianie podstawowego założenia; system dysponujący strukturami reprezentacji powinien natomiast potrafić ponownie wyprowadzić konsekwencje po modyfikacji jednego elementu modelu.

Nie jest to jednak dychotomia absolutna. W praktyce kompetencje mogą być stopniowalne, zależne od domeny i wrażliwe na sposób sformułowania problemu. Właśnie dlatego niezbędna jest cała rodzina testów, a nie pojedyncze, magiczne zadanie rozstrzygające o „prawdziwym rozumieniu”. Copycat uczy również, że impas nie musi być wyłącznie porażką. W architekturze płynnych pojęć blokada stanowi informację o nieadekwatności aktualnej interpretacji. Może ona prowadzić do zwiększenia temperatury i poszerzenia aktywacji konceptualnych otoczek, co czyni bardziej prawdopodobnymi „poślizgi”, które wcześniej były zbyt odległe.

W poznaniu ludzkim analogiczna funkcja impasu jest dobrze znana: brak postępu może skłaniać do porzucenia niejawnych ograniczeń i przedefiniowania problemu. Mechanizm ten nie zawsze jest skuteczny – człowiek równie często uporczywie powtarza błędną strategię – lecz zdolność rozpoznawania własnej nieskuteczności jest jednym z warunków adaptacyjnego rozumowania. W tym miejscu powraca dylemat eksploracji i eksploatacji. Gdy system posiada trafną hipotezę, racjonalne jest koncentrowanie zasobów na jej rozwijaniu; gdy jednak hipoteza konsekwentnie zawodzi, zasadna staje się szersza eksploracja. Copycat próbuje regulować tę proporcję wewnętrznie poprzez temperaturę i jakość powstających struktur.

Tak rozumiana inteligencja nie jest więc maksymalną deterministycznością, lecz umiejętnością właściwego zarządzania własną niepewnością. System powinien funkcjonalnie wiedzieć, kiedy ufać dotychczasowej reprezentacji, a kiedy przestać jej ufać. Zasada ta zbliża nas do problemu metaracjonalności. Racjonalność pierwszego rzędu dotyczy wyboru działania przy danych przekonaniach i celach. Metaracjonalność odnosi się natomiast do wyboru samego sposobu rozumowania: ile czasu poświęcić na poszukiwania, kiedy zmienić model, które uproszczenia są dopuszczalne i w którym momencie koszty dalszej analizy przewyższają oczekiwaną korzyść. Copycat nie stanowi rozwiniętej teorii metaracjonalności, lecz jego architektura zawiera jej elementarną intuicję. System reguluje nie tylko treść własnych hipotez, ale również intensywność i charakter eksploracji. Dlatego najważniejszym osiągnięciem projektu nie jest konkretny zestaw rozwiązanych analogii literowych, lecz próba formalizacji myśli, że rozumowanie obejmuje jednocześnie wybór reprezentacji, ocenę jej jakości oraz sterowanie procesem dalszego poszukiwania.

Na każdym z tych poziomów mogą wystąpić błędy. System może wybrać niewłaściwe elementy, błędnie zinterpretować relacje lub zbyt wcześnie ustabilizować się wokół przeciętnej hipotezy. Inteligencja nie jest zatem pojedynczą zdolnością, lecz dynamiczną koordynacją procesów, które wzajemnie kontrolują swoje wyniki.

Konfrontacja płynności FARG ze współczesnymi modelami językowymi

Właśnie w tym punkcie następuje najciekawsze spotkanie tradycji FARG ze współczesną sztuczną inteligencją. Dzisiejsze modele językowe potrafią rozwiązywać ogromną liczbę zadań analogicznych bez jawnego Slipnetu, ręcznie wpisanej sieci pojęć czy Coderacka. Fakt ten jest istotny, gdyż pokazuje, że funkcjonalne kompetencje analogiczne mogą powstawać w architekturach radykalnie innych od tych, które Hofstadter i jego współpracownicy projektowali jako modele poznania. Nie wynika z tego jednak ani że FARG została sfalsyfikowana, ani że współczesne systemy osiągnęły dokładnie to, co autor książki nazywał płynną inteligencją. Aby rozstrzygnąć znaczenie tego sukcesu, należy porównać nie nazwy komponentów, lecz zachowanie reprezentacji.

Pojawia się zatem pytanie: czy model językowy rzeczywiście potrafi zmienić sposób reprezentowania problemu, czy jedynie generuje wynik, który z zewnątrz przypomina rezultat takiej zmiany? Nie da się na to odpowiedzieć, obserwując jedynie płynność rozmowy. Konieczne są badania nad transferem kompetencji, odpornością na zmianę formatu, analogiami między odległymi domenami, kontrfaktycznością oraz internalizacją nowych reguł i reakcją na problemy celowo skonstruowane poza typowymi wzorcami treningowymi. Co więcej, trzeba oddzielić sukces behawioralny od teorii mechanizmu, który za nim stoi – ta sama odpowiedź może być bowiem efektem różnych procesów wewnętrznych. W ten sposób Copycat prowadzi nas do momentu, w którym historyczny spór między symboliczną AI, konekcjonizmem a programem FARG przestaje wystarczać. Wielkie modele językowe stworzyły klasę systemów, których kompetencji nie da się uczciwie opisać ani jako klasycznej manipulacji symbolami, ani jako prostego magazynu statystycznych skojarzeń.

Zarazem ich niezwykła sprawność nie rozwiązała problemu reprezentacji, lecz uczyniła go trudniejszym. Obserwujemy dziś zachowania przypominające abstrakcję, analogię i rekontekstualizację na skalę nieosiągalną dla dawnych mikrodomen. Znacznie mniej oczywiste jest jednak to, co dokładnie zostało wewnątrz tych systemów wyuczone, w jaki sposób reprezentowane są relacje oraz dlaczego ta sama maszyna potrafi raz dokonać subtelnej analogii, a chwilę później rozpaść się na problemie, który człowiek uznałby za banalny. Właśnie ta nierówność kompetencji będzie punktem wyjścia kolejnej fazy analizy. Zanim jednak przejdziemy bezpośrednio do modeli językowych, należy przyrzeć się, jak FARG próbował wyprowadzić płynność poza uporządkowaną domenę alfabetu. Copycat operował w świecie „krystalicznym”, gdzie elementy i podstawowe relacje były stosunkowo dobrze określone. Inteligencja codzienna działa natomiast w środowisku nieuporządkowanym, w którym sama identyfikacja obiektu, granicy sytuacji czy relewantnej relacji

jest już częścią problemu. Przejście od Copycata do Tabletop będzie zatem przejściem od analogii pomiędzy gotowymi symbolami do pytania: jak umysł wykrawa z chaotycznego świata coś, co w ogóle nadaje się do porównania?

Ostatecznie pytanie o naturę inteligencji przesuwa się z obszaru 'ile system wie' na płaszczyznę 'jak zmienia sposób organizowania tego, co wie'. Być może prawdziwym sprawdzianem umysłu nie jest podanie poprawnej odpowiedzi, lecz zdolność do przeżycia impasu i przekucia go w moment wglądu. Czy zatem maszyna, która nigdy nie błądzi w sposób ludzki, może kiedykolwiek naprawdę odkryć coś nowego?

Inspiracje

[1]



"Fluid Concepts and Creative Analogies" Douglas Hofstadter

ISBN: 9780140258356

Douglas Richard Hofstadter (ur. 15 lutego 1945 w Nowym Jorku) jest amerykańskim uczonym i wybitnym profesorem kognitywistyki i literatury porównawczej na Uniwersytecie Indiany w Bloomington. Uzyskał doktorat z fizyki na Uniwersytecie Oregon w 1975 roku, po czym zajął się kognitywistyką i sztuczną inteligencją. Hofstadter jest najbardziej znany ze swojej przełomowej, multidyscyplinarnej książki „Gödel, Escher, Bach: an Eternal Golden Braid” (1979), która zdobyła Nagrodę Pulitzera w kategorii literatury faktu oraz National Book Award. Jego badania koncentrują się na mechanizmach ludzkiego myślenia, świadomości, tworzeniu analogii i autoreferencyjności, które konceptualizuje w ramach dziwnych pętli. Kierował Centrum Badań nad Koncepcjami i Poznaniem, nadzorując rozwój kognitywnych architektur obliczeniowych, takich jak Copycat. Hofstadter wniósł również znaczący wkład w teorię przekładu, językoznawstwo i filozofię umysłu.

Douglas Richard Hofstadter (born February 15, 1945, in New York City) is an American scholar and Distinguished Professor of Cognitive Science and Comparative Literature at Indiana University Bloomington. He earned his Ph.D. in physics from the University of Oregon in 1975 before transitioning into cognitive science and artificial intelligence. Hofstadter is best known for his groundbreaking multidisciplinary book Gödel, Escher, Bach: an Eternal Golden Braid (1979), which won the Pulitzer Prize for General Nonfiction and the National Book Award. His research focuses on the mechanisms of human thought, consciousness, analogy-making, and self-reference, which he conceptualizes through the framework of strange loops. He directed the Center for Research on Concepts and Cognition, overseeing the development of cognitive computational architectures such as Copycat. Hofstadter has also made prominent contributions to translation theory, linguistics, and the philosophy of mind.

Indiana University Bloomington

Literatura uzupełniająca

Książki

Książki autora i dzieła przywołane w tekście:



[1] "Metamagical Themas"

Douglas R Hofstadter

2008 | Hachette UK | ISBN: 9780786723867



[2] "The Mind's I"

Douglas Hofstadter



[3] "Gödel, Escher, Bach"

Douglas Hofstadter

1979 | ISBN: 9780140179972

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

Douglas Hofstadter

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: d7758abf-3f34-4570-bda2-17ef6f6f5b81