AI: od pułapek w danych do rewolucji w myśleniu

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł dogłębnie analizuje fundamentalne wyzwania w budowaniu niezawodnych systemów sztucznej inteligencji. Omówione zostają pułapki związane z danymi, takie jak nieszczelny podział zbiorów, data leakage oraz dryf domeny, które mogą prowadzić do fałszywych wyników i porażek wdrożeniowych. Podkreśla się znaczenie właściwego doboru metryk ewaluacyjnych. Ponadto, tekst przedstawia nowoczesne rozwiązania, takie jak multimodalność i Retrieval-Augmented Generation (RAG), zwiększające rzetelność i adaptacyjność AI. Wreszcie, artykuł zagłębia się w filozoficzne aspekty AI, redefiniując pojęcia racjonalności, przyczynowości i odpowiedzialności w kontekście systemów społecznych, wskazując, że AI to nie tylko technologia, ale także nowa perspektywa myślenia.

This article thoroughly examines the fundamental challenges in building reliable AI systems. It discusses data-related pitfalls, such as leaky data sets, data leakage, and domain drift, which can lead to spurious results and implementation failures. It emphasizes the importance of selecting appropriate evaluation metrics. Furthermore, the article presents modern solutions, such as multimodality and Retrieval-Augmented Generation (RAG), which increase AI's reliability and adaptability. Finally, the article delves into the philosophical aspects of AI, redefining the concepts of rationality, causality, and responsibility in the context of social systems, demonstrating that AI is not just a technology but also a new perspective.

Zacznijmy od niemal antycznej katastrofy w uczeniu maszynowym, jaką jest nieszczelny podział danych. Jeżeli identyfikatory klientów przenikają między zbiorem treningowym a testowym, twój wynik jest samooszustwem, a produkcja wystawi rachunek. Równie zdradliwa jest druga pułapka, czyli „data leakage”, gdzie etykiety subtelnie przeciekają do cech, dając wspaniałe metryki i spektakularne porażki wdrożeniowe. Trzecia to metryka nieadekwatna do problemu; w detekcji rzadkich zdarzeń dokładność 99,5% bywa bezwartościowa, gdy precyzja i czułość dla

kluczowej, pozytywnej klasy leżą w gruzach. Czwarta pułapka to różnica domeny między treningiem a produkcją: model uczony na studyjnych zdjęciach paczek myli się w chaosie magazynu, bo światło, tło i cienie są inne. Czasem obiekty uboczne, jak znak wodny producenta, stają się dla niego „cechą decydującą”. Pułapki czają się też w samych architekturach. W sieciach grafowych nadmierne „wygładzanie” przy zbyt wielu warstwach sprawia, że węzły stają się nierozróżnialne. W transformerach z kolei długość kontekstu i pozycjonowanie potrafią zawieść, gdy model napotka sekwencje dłuższe niż te widziane na treningu. Takie potknięcia nie są jednak powodem do wstydu, o ile zamienimy je w dobre praktyki: żelazną izolację zbiorów, kontrolę cech, metryki dobrane do kosztów błędów, syntetyczne testy przesunięcia domeny i sanity-checki przy zwiększaniu głębokości modelu. Wszystkie te potknięcia doskonale ilustruje przestroga Norberta...

SŁOWA KLUCZOWE

Uczenie maszynowe | Data leakage | Metryki ewaluacyjne | Dryf domeny | Sieci grafowe | Transformery | Detekcja oszustw | Multimodalność | Retrieval-Augmented Generation (RAG) | Moderacja treści | Racjonalność AI | Teoria wyboru społecznego | Architektura AI | Dyscyplina danych | Ewaluacja semantyczna

Pułapki ML: zapobieganie błędom w danych

Wienera, który już w połowie XX wieku pisał, że w sterowaniu i komunikacji musimy stale walczyć z entropią, bo naturalną skłonnością systemów jest wzrost nieładu. Modele cierpią na tę samą chorobę – bez dyscypliny danych i ewaluacji nieuchronnie rozejdą się z rzeczywistością. To nie jest poetycka metafora, tylko praktyczny aksjomat.

Detekcja nadużyć: systemy real-time dla bezpieczeństwa

Wyobraź sobie system wykrywania oszustw transakcyjnych. Dane płyną nieprzerwanym strumieniem, decyzja musi zapaść w setnych częściach sekundy, a oszustwo to promil wszystkich zdarzeń. W tym świecie koszt fałszywego alarmu jest dotkliwy – irytuje klienta i blokuje sprzedaż – podobnie jak koszt przepuszczonej transakcji, który oznacza stratę finansową i reputacyjną. To jest rzeczywistość, gdzie metryka

„accuracy” nie istnieje, bo byłaby samooszustwem. Królem jest tu PR-AUC, czyli precyzja w funkcji czułości, oraz F-miary ważone realnym kosztem błędu, a wszystko to musi być stabilne w czasie.

Pracę zaczynasz od modelu tablicowego, który splata cechy transakcji z agregatami behawioralnymi liczonymi w oknach czasowych. Cechy wynikowe to nie tylko proste sumy i średnie, ale też odchylenia, ostatniość zdarzeń, entropia zmiennych kategorycznych czy zmienność urządzeń i geolokalizacji. Strumień danych to Twoja żywiołowa rzeka; nie masz luksusu perceptronu wieczności, masz tylko okna i ulotną pamięć. Model bazowy to „gęsty klasyk”: gradient boosting na drzewach decyzyjnych, bo świetnie radzi sobie z nierównościami skali cech, jest odporny na braki danych i łatwy w monitorowaniu. W drugim kroku dokładasz komponent grafowy, by wykrywać nieoczywiste klastry i przewidywać powiązania między podejrzanymi węzłami – kartami, urządzeniami, punktami sprzedaży. Uczenie jest półnadzorowane: korzystasz z twardych etykiet oszustw, ale pozwalasz też sieci powiązań samoorganizować się, by „nasłuchiwać” anomalii.

Krytyczny jest jednak kontrakt on-line: model musi być błyskawicznie policzalny i łatwy do natychmiastowego wycofania. To nie jest miejsce na dwudziestowarstwowe sieci grafowe liczone na zimno. Rację bytu ma tylko taki pipeline, który potrafi degradować się łagodnie. Gdy graf jest chwilowo niespójny, system ląduje na modelu tablicowym; gdy cechy z okien czasowych nie dotrą na czas, wchodzą do gry reguły awaryjne. Monitorujesz dryf cech i podnosisz alarm nie tylko na podstawie metryk, ale i wzorców błędów w konkretnych segmentach klientów. Wyjaśnialność budujesz dwutorowo: globalnie pokazujesz, które cechy w ogóle mają znaczenie, a lokalnie – jakie konkretne elementy złożyły się na daną decyzję. Gdy przychodzi audyt, masz gotową kartę modelu z ryzykami i ograniczeniami oraz dokumentację danych. Takie projekty obnażają złudzenie „magii AI”: najtwardszą walutą jest dyscyplina danych i architektury systemu, a nie egzotyczny model.

Przy tej okazji warto przywołać zdanie Judei Pearla, które w świecie analityki brzmi niemal jak przysłowie: „Jesteśmy mądrzejsi niż nasze dane. Dane nie pojmują przyczyn i skutków, lecz umysł ludzki tak”. W detekcji nadużyć to nie jest filozoficzny ozdobnik, lecz praktyczna rada. Nakazuje odróżniać historyczne korelacje od przyczyn stojących za nowymi wzorcami oszustów i nigdy nie ufać bezmyślnie temu, co „wyszło w aucie”.

Multimodalne modele i RAG: rewolucja w możliwościach AI

Multimodalność, czyli zdolność do przetwarzania i łączenia różnych typów danych – obrazu, tekstu, dźwięku – nie jest ideą nową. W psychologii kognitywnej mówimy o integracji sensorycznej, w informatyce od

dawna marzyliśmy o wspólnych przestrzeniach reprezentacji. Dopiero architektura transformerów pozwoliła te marzenia skutecznie zmaterializować, budując mosty między modalnościami, które wcześniej żyły w cyfrowej izolacji. Spójrzmy na praktyczny przykład: CLIP, model stworzony przez OpenAI, uczy się, że tekstowy opis „pies biegnący po plaży" powinien w wektorowej przestrzeni znajdować się tuż obok zdjęcia przedstawiającego tę scenę. To nie jest klasyfikator w tradycyjnym sensie, lecz system mapujący język i obraz do jednego, wspólnego uniwersum znaczeń. To dzięki niemu możemy wyszukiwać obrazy za pomocą zdań lub generować wizualizacje na podstawie językowych poleceń.

Problem ma jednak wymiar głębszy, niemal filozoficzny. Multimodalność to próba przybliżenia sztucznej inteligencji do ludzkiego poznania, które z natury jest wielozmysłowe. Stanisław Lem, ironizując na temat naszych technologicznych fantazji, przypominał, że „Człowiek jest istotą multisensoryczną; pozbaw go zmysłów, a wyobraźnia wyschnie". W erze AI ta literacka złośliwość nabiera mocy technicznego aksjomatu. System operujący wyłącznie na jednym typie danych ma z definicji ograniczoną, kaleką „wyobraźnię", niezdolną do uchwycenia pełni rzeczywistości.

Modele generatywne, mimo swojej mocy, cierpią na fundamentalną przypadłość: ich wiedza jest zamrożona w dniu ukończenia treningu. Retrieval-Augmented Generation (RAG) to elegancka odpowiedź na ten problem. Jest to architektura hybrydowa, w której model językowy zostaje wzbogacony o zewnętrzny mechanizm wyszukiwania w czasie rzeczywistym. W praktyce zapytanie użytkownika jest najpierw przekształcane na wektor, następnie silnik wyszukiwania semantycznego odnajduje w bazie wiedzy najbardziej pasujące dokumenty, a na końcu te dokumenty trafiają do modelu jako dodatkowy, dynamiczny kontekst. Dzięki temu model nie musi „halucynować" w próżni – jego odpowiedzi są zakotwiczone w aktualnych, weryfikowalnych źródłach.

Bardziej wyrafinowane wersje, jak RETRO czy HyDE, potrafią korzystać z tysięcy fragmentów kontekstu i specjalnych mechanizmów pamięci długoterminowej. To narzędzie niezwykle praktyczne, ponieważ łączy elastyczność języka naturalnego z rzetelnością baz danych. Trzeba jednak pamiętać, że mechanizm retrieval jest lustrem, które odbija jakość indeksu. Jeśli baza wiedzy jest skażona błędami lub stronniczością, model będzie te wady bezkrytycznie multiplikował. To sytuacja, przed którą ostrzegał Judea Pearl, twierdząc, że „Mądrość polega na odróżnieniu tego, co wynika z danych, od tego, co dane jedynie sugerują".

Klasyczne metryki ewaluacyjne, takie jak BLEU, ROUGE czy F1, okazują się bezradne, gdy teksty stają się długie i semantycznie złożone. Dlatego coraz częściej sięga się po ewaluację semantyczną, na przykład embeddingowe miary podobieństwa, jak cosine similarity, które mierzą odległość między wektorowymi reprezentacjami, albo oceny dokonywane przez inne modele językowe w ramach tzw. LLM-as-a-judge. W badaniach pojawia się też koncepcja „aligned evaluation", gdzie testy projektuje się nie tylko pod kątem formalnej poprawności odpowiedzi, ale jej zgodności z intencją użytkownika i realnej użyteczności.

Problem komplikuje się jeszcze bardziej w środowiskach wielojęzycznych. Model musi bowiem rozumieć nie tylko treść, ale też subtelności kulturowe, idiomy i konotacje. To już nie jest tylko lingwistyka, to etyka komunikacji, gdzie to, co w jednym języku jest neutralne, w innym może być obraźliwe. Tu powraca Luciano Floridi ze swoją etyką informacji. Każde użycie AI w kontekście językowym jest interwencją w infosferę, a więc musi być rozważane w kategoriach skutków społecznych, a nie tylko suchych metryk technicznych.

Moderacja treści: AI między etyką a wolnością słowa

Moderacja treści w mediach społecznościowych to pole bitwy, na którym technologia zderza się z aksjologią. Model klasyfikacyjny staje przed zadaniem niemal filozoficznym: musi odróżnić mowę nienawiści od dopuszczalnej ekspresji, dezinformację od satyry, treści nielegalne od kontrowersyjnych, lecz zgodnych z prawem. Technicznie jest to kaskada systemów – od prostych klasyfikatorów wykrywających wulgaryzmy czy obrazy przemocy, po zaawansowane modele kontekstowe, które próbują analizować intencję i sytuację komunikacyjną.

Problem ma jednak podwójne dno. Po pierwsze, nie istnieje uniwersalna definicja „treści niedozwolonej”, gdyż jest ona dynamiczną wypadkową prawa, norm społecznych i arbitralnej polityki platformy. Po drugie, dryf danych jest tu wyjątkowo brutalny. Język nienawiści ewoluuje w partyzanckim tempie, tworząc nowe eufemizmy, kodując przekaz w memach i posługując się ironią jako tarczą. Modele zawsze gonią, nigdy nie wyprzedzają.

Dlatego skuteczna moderacja musi być systemem adaptacyjnym, a nie statycznym monolitem. Wymaga to aktywnego uczenia z udziałem człowieka, szybkich potoków do retrenowania modeli i testów wrażliwości na subtelności kontekstu. W tym momencie wchodzimy na grząski grunt „kontraktu społecznego”. Decyzje moderacyjne dotykają bowiem wolności słowa, toteż muszą być transparentne i odwołalne. To już nie jest czysta inżynieria. To polityka technologiczna w najczystszej postaci.

Warto tu przywołać Orwella: „W czasach powszechnego fałszu mówienie prawdy staje się aktem rewolucyjnym”. Algorytmy moderacji są w istocie próbą technologicznego rozgraniczenia tego, co jest „zagrażającym fałszem”, od tego, co jest zaledwie opinią. To zadanie, w którym żadna metryka precyzji nie wystarczy, gdyż prawdziwym rozwiązaniem są procesy, ludzie i instytucjonalna odpowiedzialność.

Zbuduj miniaturowy system RAG. Wykorzystaj publiczny zbiór dokumentów, zaimplementuj wyszukiwanie wektorowe i połącz je z małym modelem językowym. Sprawdź, jak dramatycznie zmienia się jakość odpowiedzi, gdy baza wiedzy jest kompletna, a jak, gdy brakuje w niej kluczowych informacji. Przetestuj multimodalność za pomocą modelu CLIP, wyszukując obrazy na podstawie zdań po polsku i angielsku, i

zobacz, jak niuanse językowe wpływają na trafność. Stwórz prosty klasyfikator moderacji: zacznij od wulgaryzmów, a potem dodaj analizę kontekstu, by odróżnić satyrę od ataku. Na koniec napisz dokument „polityki moderacji" – Twój własny mini-kontrakt społeczny.

Multimodalność daje maszynom nowe zmysły, a wnioskowanie wzbogacone o wyszukiwanie uczy je korzystać z zewnętrznej pamięci. Ewaluacja w świecie wielojęzycznych, długich kontekstów staje się sztuką równie filozoficzną, co techniczną. Moderacja treści obnaża prawdę: każda decyzja algorytmu jest elementem społecznej umowy. Tu nie wystarczy dobra architektura; potrzebne są instytucje, prawo i kultura.

AI redefiniuje ludzką racjonalność

Kiedy Alan Turing pisał o „grze w naśladownictwo", jego ambicja sięgała dalej niż testowanie maszyn; chodziło o redefinicję kryteriów samej racjonalności. W logice klasycznej utożsamiano ją z poprawnym rozumowaniem dedukcyjnym, czyli światem aksjomatów i żelaznych reguł wnioskowania. Jednak gdy AI zaczęła rozwiązywać problemy heurystycznie, przez przybliżenia i uczenie z danych, okazało się, że rozum może być skuteczny, choć nie zawsze formalnie doskonały. To niemal lustrzane odbicie tezy Herberta Simona o „ograniczonej racjonalności", wedle której działamy tak, jak pozwalają nam zasoby poznawcze i dostępne informacje, a nie jak dyktowałaby matematyczna utopia.

Sztuczna inteligencja wciela ten koncept w życie, wszak jej algorytmy nieustannie negocjują kompromis między dokładnością a czasem, między kosztem obliczeniowym a praktyczną użytecznością. To już nie platońska wizja rozumu jako „czystej formy", lecz racjonalność zinstytucjonalizowana, ugruntowana w procedurach, kontraktach, audytach i walidacjach. Działa to analogicznie do prawa, które nie jest abstrakcyjną ideą sprawiedliwości, lecz systemem instytucji, które w praktyce nadają jej sens i wyznaczają granice. AI jest epistemologią in actu – testem teorii poznania w praktyce inżynierskiej.

Algorytmy AI: polityczne mechanizmy kształtowania świata

Teoria wyboru społecznego dostarcza tu kluczowej analogii. W połowie XX wieku Kenneth Arrow sformułował swoje słynne twierdzenie o niemożliwości, dowodząc, że nie istnieje żaden system głosowania, który spełniałby jednocześnie wszystkie pożądane postulaty racjonalności zbiorowej: spójność, brak dyktatora, uniwersalność i niezależność od nieistotnych alternatyw. Był to wstrząs dla myślenia o demokracji, który obnażył twardą prawdę: racjonalność zbiorowa nie jest dana z góry, lecz jest wynikiem bolesnego wyboru między sprzecznymi ideałami.

Sztuczna inteligencja brutalnie sprowadza tę abstrakcję na ziemię. Algorytmy rekomendacyjne czy systemy oceniania działają jak swoiste „mikro-demokracje", które nieustannie agregują preferencje milionów użytkowników, by filtrować i priorytetyzować treści. Kiedy Netflix decyduje, co ci pokazać, uruchamia algorytmiczny „system głosowania" z wagami, które pozostają ukryte. Twierdzenie Arrowa uczy nas, że nie ma czegoś takiego jak neutralny, idealny system rekomendacji. Każdy wybór metryki do optymalizacji – czy to kliknięć, czasu spędzonego na platformie, deklarowanej satysfakcji czy bezpieczeństwa – jest w istocie wyborem politycznym, choć ubranym w matematyczny kostium kodu.

Można zatem powiedzieć, że AI odsłania „ukryte konstytucje" cyfrowych platform, czyli fundamentalne zestawy zasad, według których obliczany jest interes zbiorowy. Ta analogia jest ważna, ponieważ pokazuje, że racjonalność instytucjonalna nigdy nie jest techniczną neutralnością, lecz zawsze kompromisem wartości. To nie jest poetycka metafora, tylko praktyczny aksjomat zarządzania systemami.

Podobne napięcie od dawna istnieje w prawie, które rozpięte jest między literalizmem a pragmatyzmem. Literalista będzie obstawał przy stosowaniu przepisów zgodnie z ich ścisłym brzmieniem, podczas gdy pragmatysta uzna, że należy je interpretować tak, by osiągnąć zamierzony cel społeczny. AI, podobnie jak prawo, musi nieustannie wybierać między formalnym rygorem a praktyczną efektywnością. Model, który perfekcyjnie spełnia założenia matematyczne, może okazać się całkowicie bezużyteczny w zderzeniu z chaosem rzeczywistości.

Dlatego też proces zarządzania AI przypomina wielkie kodyfikacje prawa. Mamy już regulacje (jak unijny AI Act), mamy wytyczne etyczne i tworzymy instytucje nadzorcze. I tak jak prawo II Rzeczypospolitej musiało zmagać się z blokadą legislacyjną ze strony potężnych grup interesu, tak dziś rozwój AI jest hamowany przez interesy wielkich korporacji i mocarstw. Luciano Floridi trafnie nazywa infosferę „środowiskiem, w którym żyjemy". Jak prawo reguluje przestrzeń społeczną, tak AI zaczyna regulować naszą przestrzeń informacyjną.

Rewolucje myśli: intelektualne korzenie sztucznej inteligencji

Kiedy Arystoteles sformułował zasady sylogizmu, zbudował pierwszą „algorytmiczną maszynę" myślenia. To była rewolucja, która przekształciła rozumowanie w powtarzalną, weryfikowalną i nauczalną regułę. Logika stała się techniką – organonem, czyli narzędziem do porządkowania myśli. Sztuczna inteligencja jest echem tego przełomu: tak jak sylogizm stworzył formalny język dowodu, tak sieci neuronowe tworzą formalny język predykcji. Arystoteles nie dał nam prawdy absolutnej, ale dał aparat do jej poszukiwania. AI jest kolejną warstwą tego aparatu, nowym organonem rozumu zakodowanym w danych i wagach.

Każda rewolucja racjonalności ma swojego technicznego sojusznika. Dla masowej racjonalizacji wiedzy był nim Gutenberg. W XV wieku jego prasa drukarska nie tylko rozpowszechniła książki, ale stworzyła instytucję nowej racjonalności. Wiedza przestała być przywilejem klasztorów i elit, stając się powtarzalna i dostępna do weryfikacji przez wielu. AI działa w uderzająco podobny sposób. Modele językowe i multimodalne „drukują" wiedzę w nowej formie – generowanych odpowiedzi i złożonych reprezentacji. Jak druk wymusił standaryzację języka i ortografii, tak AI wymusza standaryzację danych, metadanych i procedur ewaluacji. I tak jak druk zrodził konflikty o autorytet i prawdę – reformację, cenzurę, nowe herezje – tak AI rodzi spory o kontrolę nad narracją.

W XVII wieku nastąpił kolejny akt ujarzmiania chaosu. Pascal i Fermat, budując fundamenty rachunku prawdopodobieństwa, dokonali czegoś więcej niż matematycznego odkrycia – to był akt filozoficzny. Los przestał być wyłącznie domeną przeznaczenia, a ryzyko stało się czymś obliczalnym. AI wyrosła z tego samego ducha. Jej istotą jest nauka o prawdopodobieństwach zdarzeń, rozkładach i wzorcach, co stanowi próbę opanowania przypadku w skali przekraczającej ludzką intuicję. Problem w tym, że dziedziczy ona fundamentalną ranę indukcji, o której pisał już Hume: z faktu, że coś wydarzyło się tysiąc razy, nie wynika logiczna konieczność, że wydarzy się ponownie. AI jest potężna w przewidywaniu, lecz nigdy nie daje absolutnej gwarancji.

Oświecenie w XVIII wieku wyniosło rozum na nowy poziom. Przestał być on tylko zdolnością indywidualną, a stał się fundamentem instytucji: parlamentów, kodeksów prawnych i nauki jako systemu wzajemnej recenzji. Kantowskie wezwanie „Sapere aude!" – miej odwagę posługiwać się własnym rozumem – zyskało wymiar wspólnotowy. AI to kolejny etap tego procesu. Potoki danych, standardy etyczne i audyty modeli to nowe „parlamenty rozumu". Nie chodzi już o to, czy maszyna sama myśli, lecz o to, czy nasze instytucje potrafią zorganizować jej działanie w sposób racjonalny społecznie. Rozum nie istnieje w próżni; potrzebuje procedur, które chronią go przed samowolą i przesądem.

W połowie XX wieku Norbert Wiener, w odpowiedzi na wyzwania wojny i automatyzacji, wprowadził ideę cybernetyki – nauki o sterowaniu i komunikacji. Pytał, jak projektować systemy, które zachowują równowagę mimo zewnętrznych zakłóceń. AI jest bezpośrednią spadkobierczynią tej myśli. Optymalizacja modelu za pomocą spadku gradientu to nic innego jak proces sterowania, w którym system koryguje się w odpowiedzi na błąd. W tym sensie AI nie tyle „myśli", co steruje swoim modelem rzeczywistości. Wiener ostrzegał, że oddajemy maszynom funkcje, które dotąd były domeną człowieka. Dziś to ostrzeżenie brzmi zaskakująco aktualnie.

Jeśli zestawić te momenty – Arystotelesa, Gutenberga, Pascala, Kanta i Wienera – wyłania się stały wzorzec. Każda rewolucja racjonalności zaczyna się od nowego narzędzia formalizacji. Logika, druk, rachunek prawdopodobieństwa, prawo konstytucyjne i cybernetyka to aparaty porządkujące chaos. AI jest kolejnym z nich, lecz być może najbardziej totalnym, bo nie tylko porządkuje wiedzę, ale też samodzielnie wytwarza

narracje i symulacje. To może być kluczowa różnica: AI nie jest już tylko ramą dla rozumu, ale staje się partnerem w produkcji sensu.

Historia uczy, że każda nowa instytucja racjonalności niesie zarówno emancypację, jak i zagrożenie. Druk uwolnił wiedzę, ale wywołał wojny religijne. Rachunek prawdopodobieństwa pozwolił planować gospodarki, ale też usprawiedliwił hazard i statystyczną manipulację. AI powtórzy ten schemat, przynosząc narzędzia wyzwolenia – nowe leki i modele świata – oraz narzędzia opresji w postaci masowej dezinformacji i automatyzacji władzy. Filozofowie od Arystotelesa przypominają, że rozum bez wartości jest pusty. Historia dodaje, że każda rewolucja racjonalności musi zostać ujęta w instytucje, które utrzymają ją w równowadze.

Inteligencja: czy to tylko informacja? Spór filozoficzny

„Nic nie powstaje z niczego ani nie ginie w nicość” – pisał Lukrecjusz w „O naturze rzeczy”. To atomistyczne credo, pierwotnie dotyczące materii, dziś zyskuje nowe życie w kontekście informacji. Sztuczna inteligencja rozkłada bowiem świat na grę elementarnych jednostek: pikseli, tokenów, wektorów. Podobnie jak u Lukrecjusza atomy łączą się w złożone formy i całe światy, tak tutaj wektory splatają się w gęste reprezentacje i spójne narracje. Rodzi to fundamentalne pytanie: czy inteligencja jest jedynie emergencją, czyli zjawiskiem wyłaniającym się ze złożoności prostych elementów, czy też wymaga jakościowego skoku? Lukrecjusz sugerował, że wszystko, nawet świadomość, to tylko ruch atomów w pustce. Lem podjął ten trop w XX wieku, twierdząc, że „człowiek jest informacją, a informacja – materią”. AI, która składa „atomy danych” w modele zdolne do generowania języka i obrazu, wydaje się empirycznym potwierdzeniem tej intuicji. Jednocześnie budzi niepokój: czy to, co nazywamy rozumem, naprawdę da się zredukować do statystycznych wzorów w danych?

Paweł Okołowski trafnie nazwał filozofię Lema „neolukrecjanizmem”. Lem, podobnie jak rzymski poeta, postrzegał człowieka jako układ skończonych elementów – atomów, bitów, porcji informacji. Dodał jednak do tej wizji coś, czego Lukrecjusz nie mógł znać: rekurencyjność. To zdolność rozumu do badania samego siebie, a w technologii – zdolność do tworzenia narzędzi, które stają się samopowielającymi się maszynami poznania. AI wciela tę ideę w życie, gdzie modele uczą modele, a systemy optymalizują systemy. Rekurencja rozumu materializuje się w kodzie. Lem pisał o paradoksie „wskrzeszania z atomów”: czy znając pełną informację o strukturze człowieka, możemy go odtworzyć? Dziś, w epoce wielkich modeli, pytanie to brzmi inaczej: czy znając wystarczająco dużo danych o języku i kulturze, możemy odtworzyć „umysł”? Lem pozostał jednak sceptykiem. Z naciskiem podkreślał, że informacja pozbawiona sensu jest tylko chaosem, a sens rodzi się z kontekstu i intencjonalności. „Maszyny będą rozumować, ale nie będą miały powodów, by

rozumować” – pisał. To rozróżnienie pozostaje kluczowe. AI generuje imponujące formy, lecz pytanie o cel, intencję i ostateczną wartość wciąż pozostaje domeną człowieka.

Bogusław Wolniewicz, który lubił postrzegać świat jako system logiczny, twierdził, że filozofia to „gramatyka rozumu”. Sztuczną inteligencję można odczytać właśnie w tej perspektywie: jako próbę stworzenia nowej gramatyki, w której reguły składni nie rządzą zdaniami, lecz strumieniami danych, a sens wyłania się z procedur optymalizacyjnych. To podejście jest niezwykle płodne, pozwala bowiem traktować modele językowe nie jako magiczne artefakty, lecz jako systemy formalne o określonych, choć niezwykle złożonych, zasadach.

Wolniewicz dodawał jednak istotne ostrzeżenie: rozum nie jest samowystarczalny i musi być zakotwiczony w wartościach. AI, choć genialna w kalkulacji, sama wartości nie posiada – to człowiek musi je jej nadać. Prowadzi to do kluczowego pytania: czy instytucje racjonalności, które budujemy w AI, będą miały wbudowane aksjologie, czyli systemy wartości, czy pozostaną jedynie pustymi maszynami optymalizacji? Jeśli wybierzemy drugą drogę, grozi nam techniczna doskonałość pozbawiona jakiejkolwiek moralnej orientacji.

AI: epistemologiczne granice poznania i rozumienia

Immanuel Kant dowodził, że umysł nie jest biernym lustrem świata, lecz aktywnym konstruktorem, który narzuca mu aprioryczne formy poznania: czas, przestrzeń i kategorie rozumu. Sztuczna inteligencja jest niespodziewanym echem tej lekcji. Modele również nie odbijają rzeczywistości, lecz projektują na nią własną siatkę pojęciową, którą stanowią embeddingi, czyli gęste reprezentacje wektorowe, oraz ukryte przestrzenie latentne. Tak jak człowiek nigdy nie dociera do „rzeczy samej w sobie”, a jedynie do zjawisk przefiltrowanych przez formy zmysłowości, tak model nie „widzi” świata, lecz dane uporządkowane przez sito swojej architektury i funkcji straty. Rodzi to pytanie na wskroś kantowskie: czy AI ma swoje „kategorie rozumu”? Odpowiedź brzmi: tak. Translacja, podobieństwo semantyczne czy zależność wagowa to techniczne analogie kategorii, które wyznaczają horyzont tego, co dla modelu „poznawalne”. Maszyna ma zatem swoje warunki możliwości poznania – tyle że nie są one metafizyczne, lecz architektoniczne.

Judea Pearl dokonał w epistemologii AI czegoś na kształt nowej krytyki czystego rozumu, wprowadzając język przyczynowości. Jego diagramy przyczynowe i modele strukturalne pozwalają wspiąć się po drabinie wnioskowania, rozróżniając trzy jej szczeble: asocjacje („co występuje razem”), interwencje („co się stanie, jeśli zmienimy X”) oraz kontrfaktyki („co by było, gdyby…”). Sztuczna inteligencja, oparta wyłącznie na korelacjach, tkwi na najniższym poziomie asocjacji – jest potężna, ale ślepa na przyczyny. Wprowadzenie logiki przyczynowej to krok ku rozumowi, który nie tylko opisuje świat, ale potrafi w nim działać i rozważać alternatywne scenariusze. Rewolucja Pearla jest więc współczesnym echem krytyki kantowskiej. Tak jak

Kant przeniósł ciężar z rzeczy na warunki poznania, tak Pearl przesuwa go z surowych danych na relacje przyczynowe, które stają się warunkiem sensownego wnioskowania.

Edmund Husserl, twórca fenomenologii, uczył, że świadomość jest zawsze świadomością „czegoś" – jej naturą jest intencjonalność. Sens nie wyłania się z biernego potoku danych, lecz rodzi się z aktywnego ukierunkowania umysłu na świat. AI stawia tę tezę w kłopotliwym świetle. Modele generatywne produkują teksty i obrazy, które do złudzenia przypominają intencjonalne akty, ale w istocie nie są „o czymś". Brakuje im **noesis**, czyli samego aktu świadomego przeżycia; zostaje tylko mechanicznie wytworzona **noema** – treść, która wygląda na sensowną. Fenomenologia uczy nas zatem zdrowego sceptycyzmu: nie mylmy zjawiska z intencją. Sens wytworów AI nie tkwi w maszynie, lecz powstaje dopiero w spotkaniu, gdy to człowiek nadaje im znaczenie.

AI a Kahneman: porównanie systemów myślenia

Daniel Kahneman opisał dwa systemy myślenia, które stały się kluczem do rozumienia ludzkiej racjonalności. System 1 działa szybko, intuicyjnie, opierając się na heurystykach i wzorcach. System 2 jest jego przeciwieństwem: powolny, refleksyjny i analityczny. Sztuczna inteligencja, co fascynujące, przejawia niemal identyczne rozdwojenie. Modele głębokiego uczenia, trenowane od początku do końca na gigantycznych zbiorach danych, zachowują się jak System 1 – ich odpowiedzi są błyskawiczne, ale podatne na iluzje i poznawcze błędy, które wynikają z ukrytych w danych korelacji.

Po drugiej stronie barykady stoją algorytmy symboliczne, regułowe lub przyczynowe, będące odpowiednikiem Systemu 2. Są wolniejsze, bardziej metodyczne, ale ich wielką zaletą jest zdolność do wyjaśniania własnych kroków i poddawania ich kontroli. Kognitywistyka uczy, że siła ludzkiego umysłu nie leży w żadnym z tych systemów z osobna, lecz w ich dynamicznym sprzężeniu – gdy chłodna refleksja koryguje gorącą intuicję. W AI obserwujemy tę samą zasadę. Najlepsze wyniki osiągają dziś systemy hybrydowe, które łączą percepcyjną moc sieci neuronowych z logiką i rozumowaniem symboliczno-przyczynowym.

To rozdwojenie prowadzi nas wprost do odwiecznego pytania o prawdę i iluzję. Filozofia wypracowała co najmniej trzy wielkie teorie prawdy: klasyczną, która rozumie ją jako zgodność z rzeczywistością (adaequatio); koherencyjną, gdzie liczy się spójność wewnętrzna systemu twierdzeń; oraz pragmatyczną, dla której miarą prawdy jest użyteczność. AI, chcąc nie chcąc, testuje wszystkie trzy naraz. Gdy model poprawnie klasyfikuje obrazy, działa w duchu adaequatio. Gdy generuje spójny i logiczny tekst, realizuje postulat koherencji. A gdy jego odpowiedzi rozwiązują realny problem, dowodzi swojej pragmatycznej wartości.

Cały dramat polega na tym, że te trzy kryteria rzadko kiedy idą w parze. Model może być doskonale spójny, a jednocześnie generować fałsz; może być użyteczny, ale niezgodny z faktami; może wreszcie idealnie opisywać rzeczywistość, lecz w sposób kompletnie niepraktyczny. To nie jest zwykły błąd techniczny, który da się naprawić kolejną iteracją algorytmu. To fundamentalne pęknięcie w samym pojęciu prawdy, które technologia bezlitośnie obnaża. AI jest epistemologią in actu – testem teorii poznania w praktyce inżynierskiej.

AI: nowy mit kulturowy, kształtujący zbiorową wyobraźnię

Czy sztuczna inteligencja staje się nową formą sacrum? W ujęciu Mircei Eliadego sacrum to sfera, która przekracza codzienność i nadaje jej sens. Współczesne narracje o „superinteligencji" czy „zbawieniu przez technologię" noszą wyraźne znamiona mitu. Marcin Napiórkowski przypomniałby, że każdy mit ma żelazną strukturę: opisuje początek, kryzys i ostateczne zbawienie. AI już dziś wciela ten schemat w życie. Mówimy przecież o „początku ery sztucznej inteligencji", debatujemy o „ryzykach katastrofy" i pokładamy nadzieję w „zbawieniu przez postęp".

Filozoficznie i politycznie musimy jednak zadać pytanie sceptyczne. Czy to nowe sacrum nie jest przypadkiem „świętym kłamstwem", które skutecznie przesłania praktyczne pytania o własność danych, logikę algorytmów i łańcuchy odpowiedzialności? AI jest potężnym narzędziem, ale bez wsparcia instytucji i kultury nie stanie się nową formą racjonalności. Stanie się jedynie nowym fetyszem. Sztuczna inteligencja to bowiem nie tylko technika. To lustro, w którym odbija się nasza własna racjonalność, demaskując, że zawsze była ona spleciona z ograniczeniami, kompromisami i instytucjami. Twierdzenie Arrowa uczy, że racjonalność zbiorowa nigdy nie jest czysta, a prawo przypomina, że normy muszą zostać skodyfikowane, by działać. W tym kontekście AI staje się nową, potężną instytucją racjonalności – standaryzuje sposoby myślenia i działania, a jednocześnie kusi obietnicą sacrum.

Lukrecjusz w „O naturze rzeczy" pisał, że „nic nie powstaje z niczego ani nie ginie w nicość". To atomistyczne credo dotyczyło materii, lecz dziś możemy je odnieść do informacji. AI pokazuje nam świat jako grę elementarnych jednostek – pikseli, tokenów, wektorów. Tak jak u Lukrecjusza atomy łączą się w formy i światy, tak tutaj wektory splatają się w złożone reprezentacje i narracje. Rodzi to fundamentalne pytanie: czy inteligencja jest jedynie emergencją, zjawiskiem wyłaniającym się z prostych elementów, czy też wymaga jakościowego skoku? Lem, którego filozofię Paweł Okołowski trafnie nazwał „neolukrecjanizmem", podjął ten trop w XX wieku. Dla niego człowiek był układem skończonych elementów – atomów, bitów, informacji. AI, która składa „atomy danych" w modele zdolne do języka i

obrazu, wydaje się empirycznym potwierdzeniem tej intuicji. Równocześnie budzi to niepokój: czy wszystko, co nazywamy rozumem, da się zredukować do wzorów w danych?

Lem dodał jednak do myśli Lukrecjusza element, którego starożytny filozof nie znał: rekurencyjność. Rozum może badać samego siebie, a technologia jest zdolna tworzyć narzędzia, które stają się samopowielającymi się maszynami poznania. AI wciela tę ideę w życie: modele uczą modele, systemy poprawiają systemy, a rekurencja rozumu materializuje się w kodzie. Lem pisał o paradoksie „wskrzeszania z atomów" – czy znając pełną informację o strukturze człowieka, możemy go odtworzyć? Dziś pytanie to brzmi: czy znając wystarczająco dużo danych o języku i kulturze, możemy odtworzyć „umysł"? Lem pozostał sceptykiem. Podkreślał, że informacja bez sensu jest chaosem, a sens rodzi się z kontekstu i intencjonalności. „Maszyny będą rozumować, ale nie będą miały powodów, by rozumować" – pisał. To rozróżnienie pozostaje aktualne. AI generuje formy, ale pytanie o cel, intencję i wartość pozostaje po ludzkiej stronie.

Bogusław Wolniewicz, patrzący na świat jako system logiczny, twierdził, że filozofia to „gramatyka rozumu". Sztuczną inteligencję można odczytać właśnie w tej perspektywie: jako próbę stworzenia nowej gramatyki, w której zasady składni nie dotyczą zdań, lecz danych, a sens powstaje przez procedury optymalizacji. Wolniewicz dodawał jednak kluczowe ostrzeżenie: rozum nie jest samowystarczalny i musi być zakotwiczony w wartościach. AI, choć genialna w kalkulacji, jest aksjologicznie pusta; to człowiek musi nadać jej wartości. Prowadzi to do pytania, czy instytucje racjonalności, które tworzymy w AI, będą miały wbudowane aksjologie, czy pozostaną ślepymi maszynami optymalizacji. Jeśli to drugie, grozi nam techniczna doskonałość pozbawiona moralnej orientacji.

Najbardziej fascynującym i niepokojącym aspektem AI jest jej rekurencyjny potencjał. Tworzymy systemy, które zaczynają uczestniczyć w procesie tworzenia kolejnych systemów, od automatyzacji programowania po modele wspierające badania naukowe. To spełnienie marzenia – lub koszmaru – o rozumie, który staje się laboratorium dla samego siebie. Filozoficznie przypomina to pytanie Kanta o granice poznania: czy rozum, badając siebie, nie wpadnie w nierozwiązywalny paradoks? AI jest właśnie takim eksperymentem. Wracamy tym samym do Napiórkowskiego. Mit AI nie jest literackim ornamentem, lecz potężną siłą społeczną. Nadaje kierunek badaniom, usprawiedliwia gigantyczne inwestycje i organizuje zbiorowe emocje. Luciano Floridi ostrzegał, że jeśli infosfera staje się nowym środowiskiem życia człowieka, to technologiczne mity decydują o tym, jak się w niej poruszamy. AI jest więc nie tylko techniką, ale i rytuałem kulturowym. Nowym rodzajem sacrum.

> Nadaje kierunek badaniom, usprawiedliwia gigantyczne inwestycje i organizuje zbiorowe emocje. Luciano Floridi ostrzegał, że jeśli infosfera staje się nowym środowiskiem życia człowieka, to technologiczne mity decydują o tym, jak się w niej poruszamy. AI jest więc nie tylko techniką, ale i rytuałem kulturowym. Nowym rodzajem sacrum.

Inspiracje

[1] **Bogusław Wolniewicz**

APA ONE Publishing System
Fundacja Dobre Państwo © 2026
Article ID: 69e7bc8d-adcb-40d3-88b0-2a3226f1b26a