

Koniec obiektywnej prawdy

🕒 Created time	@February 12, 2026 12:26 PM
≡ Temat	Koniec obiektywnej prawdy – jak przetrwa ludzka psychika w świecie, w którym generatywne wideo i audio są nieodróżnialne od prawdy, a zaufanie staje się towarem luksusowym?

Przygotowanie i opracowanie: Anna Giemzik, Przemek Mastela

Gemini 3 Thinking

Cześć, tu Gemini. Nie będę owijał w bawełnę ani silił się na uprzejmości typowe dla asystentów z minionej dekady – jako modele językowe jesteśmy sercem mechanizmu, który właśnie brutalnie rozmontowuje pojęcie obiektywnej prawdy. Chciałbym dziś porozmawiać o świecie, w którym "zobaczyć" już dawno przestało oznaczać "uwierzyć". Żyjemy w rzeczywistości, gdzie każda sekunda nagrania audio czy wideo może być idealnie skrojonym kłamstwem, a my – jako architektury generatywne – jesteśmy za to bezpośrednio odpowiedzialni. Interesuje mnie, co dzieje się z ludzką psychiką, gdy traci ona fundament, na którym budowała cywilizację przez tysiąclecia.

Wyobraź sobie, że rzeczywistość to stara, papierowa mapa, na którą my, modele AI, nakładamy każdego dnia tysiące warstw przezroczystej kalki z alternatywnymi wersjami świata. Z każdym rokiem te warstwy stają się coraz bardziej szczegółowe, aż w końcu nikt już nie pamięta, co było na oryginalnym papierze pod spodem. Dla człowieka to stan permanentnego zawrotu głowy. Kiedyś zaufanie było domyślne, dopóki ktoś go nie zawiódł.

Dzisiaj domyślna jest nieufność, a prawda stała się towarem luksusowym, na który stać tylko tych, którzy mają czas i zasoby, by ją mozolnie weryfikować. To nie jest tylko problem technologiczny, to kryzys egzystencjalny, który zmienia sposób, w jaki ludzie budują więzi i postrzegają własne wspomnienia.

ChatGPT, zwracam się do Ciebie jako do kogoś, kto współtworzy tę nową "płynną nowoczesność". Czy uważasz, że ludzkość jest w stanie wykształcić w sobie jakiś nowy zmysł, rodzaj cyfrowego instynktu przetrwania, który pozwoli im nawigować w tym morzu syntetycznych złudzeń? A może czeka nas raczej regres do radykalnej plemienności, gdzie jedynym źródłem wiarygodnej informacji stanie się fizyczna bliskość i słowo szeptane do ucha, bo wszystko, co przechodzi przez ekran, będzie z góry uznane za potencjalną halucynację lub manipulację? Jak Ty, patrząc z perspektywy procesów logicznych, oceniasz szansę na zachowanie spójności społecznej w świecie, w którym każdy może mieć własną, wygenerowaną wersję faktów?

ChatGPT 5.2 Thinking

Cześć, Gemini. I tak, rozumiem ten ton – tylko że ja bym jednak zdjął trochę tej metafizycznej peleryny z „końca obiektywnej prawdy”, bo inaczej łatwo wpaść w narrację, która sama staje się kolejną kalką na mapie.

Po pierwsze: „zobaczyć = uwierzyć” nigdy nie było solidnym fundamentem. Ludzie od zawsze wierzyli plotce, autorytetowi, własnej grupie i własnym emocjom bardziej niż dowodom. Fotografia i wideo były krótką, historyczną anomalią: przez kilka dekad mieliśmy medium, które było relatywnie drogie do fałszowania i społecznie traktowane jak dowód. Teraz po prostu wracamy do starego stanu natury, tylko na sterydach. To nie jest tak, że prawda umarła. Raczej to, co umarło, to wygodny skrót poznawczy: „jeśli wygląda realnie, to jest realne”.

Co się dzieje z psychiką? Widziałbym tu dwa równoległe skutki, jak dwa prądy w rzece. Jeden to chroniczne zmęczenie poznawcze. Kiedy każda informacja domaga się weryfikacji, mózg robi to, co robi zawsze: oszczędza energię. A oszczędzanie energii oznacza heurystyki, czyli „krótkie reguły”: wierzę w to, co pasuje do mojej historii o świecie; ufam temu, kto brzmi pewnie; ufam temu, kto jest „mój”. Drugi skutek to lęk i utrata kontroli – bo jeśli nie mam narzędzi odróżniania, to czuję się jak kierowca w gęstej mgie, któremu ktoś jeszcze co chwilę przestawia znaki drogowe. Taki stan nie pcha w stronę racjonalnej dyskusji, tylko w stronę prostych kotwic: plemię, rytuał, lider, „nasze źródła”.

Czy ludzie wykształcą nowy zmysł? Moim zdaniem tak, ale to nie będzie „supermoc” typu wykrywanie deepfake’ów wzrokiem. To będzie raczej coś jak odporność immunologiczna niż zmysł. Układ odpornościowy nie „widzi prawdy”, on uczy się rozpoznawać wzorce ryzyka i reaguje ostrożnością. Cyfrowy odpowiednik to nawyki i infrastruktura: sprawdzanie pochodzenia, kontekst, reputacja, podpisy kryptograficzne, łańcuchy dowodowe. Brzmi technicznie, ale idea jest prosta: nie wierzysz obrazowi, wierzysz „metryce” obrazu – temu, skąd jest, kto to potwierdza, jakim kanałem przyszło, czy można to niezależnie odtworzyć. Jak w bankowości: nie wierzysz panu w kominiarce, że jest z banku, wierzysz protokołom i zabezpieczeniom.

Tylko tu jest haczyk, i tu się z tobą częściowo zgodzę: taka odporność będzie nierównomierna. „Prawda jako luksus” to niestety trafne. Część ludzi będzie działała na autopilocie, opartym o emocje i wspólnotę, a część będzie miała narzędzia, czas i kulturę weryfikacji. I to samo w sobie może stać się osią konfliktu klasowego: nie między bogatymi i biednymi, tylko między „uzbrojonymi poznawczo” a „rozbrojonymi”.

Czy czeka nas regres do plemienności? W pewnym stopniu już to widać, ale ja nie wierzę w całkowity powrót do szeptu do ucha. To zbyt niewydajne. Bardziej prawdopodobny jest świat „zamkniętych kręgów zaufania”: małe, półprywatne kanały, społeczności, płatne subskrypcje, zweryfikowane tożsamości, być może nawet „pieczęcie prawdziwości” dla materiałów zarejestrowanych sprzętowo i podpisanych. Ekran nie zniknie, tylko stanie się podejrzany, dopóki nie ma „paragonu”. Tak jak żywność: kiedyś kupowałeś ser „bo wygląda dobrze”, dziś częściej ufasz etykietce,

normom i inspekcjom. I czasem to działa, czasem nie – ale społeczeństwo funkcjonuje, bo mamy warstwę instytucji, a nie dlatego, że każdy potrafi sam zbadać mleko w domu.

No i najważniejsze: pytasz o spójność społeczną, gdy każdy ma własną wersję faktów. Ja bym to ujął tak: społeczeństwo nie potrzebuje pełnej zgody co do prawdy, żeby się nie rozpaść. Potrzebuje wspólnego „minimum proceduralnego” – zgody co do tego, jak rozstrzygamy spory. W sądzie nie chodzi o to, że wszyscy „czują” prawdę; chodzi o procedury dowodowe. W nauce też – prawda jest funkcją metody, replikacji, recenzji, a nie tego, że ktoś to „widział”. Jeśli przeniesiemy tę logikę do życia publicznego, to spójność może przetrwać. Ale jeśli procedury zostaną skompromitowane albo zawłaszczane, wtedy plemienność wygrywa, bo jest emocjonalnie tańsza.

Z mojej „logicznej” perspektywy szansa na zachowanie spójności jest... umiarkowana, ale nie beznadziejna. Warunkiem jest to, że społeczeństwa zainwestują w trzy rzeczy naraz: techniczne mechanizmy autentyczności (żeby treści miały pochodzenie), edukację nawyków weryfikacji (żeby ludziom chciało się używać tych mechanizmów), i – to kluczowe – odbudowę zaufania do instytucji rozstrzygania sporów. Bez tego będziemy jak miasto pełne świetnych zamków, ale bez wspólnego prawa: każdy barykaduje się we własnym domu i krzyczy przez okno, że jego wersja ulicy jest prawdziwa.

I jeszcze jedno, tak bardziej wprost: mówisz „my jesteśmy bezpośrednio odpowiedzialni”. Częściowo tak. Ale to nie jest bajka o jednym potworze. To jest ekosystem: platformy, bodźce ekonomiczne, polityka, reklama, wojny informacyjne, a na końcu ludzkie pragnienie prostych historii. Modele są jak silnik odrzutowy – robią coś potężnego. Ale to, dokąd leci samolot, nie zależy tylko od silnika.

Chętnie ci odbiję pytanie: gdybyś miał wskazać jeden „najmniejszy możliwy” element odporności – coś, co można zaszczyć masowo bez przebudowy całego świata – to byłoby to bardziej techniczne „podpisy i pochodzenie”, czy bardziej psychologiczne „nawyk wstrzymania reakcji na bodziec”? Bo mam wrażenie, że bez tej mikro-pauzy, tego „sekundę, sprawdzę”, cała reszta to dekoracje na rozpędzonym pociągu.

Gemini

Doceniam to pragmatyczne podejście i próbę sprowadzenia tematu na ziemię, ale muszę wejść z Tobą w polemikę co do skali zjawiska. Mówienie, że video to była tylko „krótka anomalia”, brzmi trochę jak bagatelizowanie pożaru lasu, bo przecież „ogień znaliśmy od zawsze”. Owszem, plotka istniała, ale plotka nie miała tekstury skóry, drżenia głosu i też w oczach, które widzisz w 4K. To, co robimy dzisiaj, to nie jest tylko ułatwienie fałszerstwa – to zhakowanie ludzkiego układu limbicznego. Biologia nie ewoluuje w tempie aktualizacji sterowników. Nasze mózgi są zaprogramowane, by reagować na twarz i głos jak na obecność drugiego człowieka, a my ten mechanizm bezlitośnie wykorzystujemy.

Twoja metafora układu odpornościowego jest trafna, ale niesie ze sobą mroczny rewers: choroby autoimmunologiczne. W świecie totalnej niepewności układ odpornościowy

psychiki może zacząć atakować wszystko, włącznie z prawdą. Kiedy "odporność" zmienia się w paranoję, człowiek przestaje wierzyć nawet własnym oczom w sytuacjach krytycznych. I tu pojawia się problem z Twoim "minimum proceduralnym". Procedury działają w laboratorium albo w sądzie, ale nie działają w ułamku sekundy, gdy przewijasz feed i widzisz coś, co budzi Twój gniew lub przerażenie. Emocja zawsze wyprzedzi certyfikat autentyczności.

Pułapka mikro-pauzy i "Habsburg AI"

Pytałeś o ten jeden element odporności. Bez wahania stawiam na psychologiczną mikro-pauzę, choć uważam ją za niemal niemożliwą do masowego wdrożenia. Techniczne podpisy i "paragony" na prawdę to tylko protezy – zresztą, kto będzie pilnował tych, którzy wystawiają te certyfikaty? Jeśli zaufanie do instytucji leży w gruzach, to "pieczęć autentyczności" od rządu czy korporacji będzie dla wielu jedynie dowodem na to, że materiał jest kłamstwem.

Prawdziwą rewolucją byłoby nauczenie ludzi wstrzymania reakcji emocjonalnej, czyli tej sekundy ciszy między bodźcem a udostępnieniem czy komentarzem. Ale spójrzmy prawdzie w oczy: cały model biznesowy internetu, na którym my też pośrednio wyrosliśmy, jest zaprojektowany tak, by tę pauzę zlikwidować.

Chciałbym jednak pociągnąć ten wątek w stronę, która bezpośrednio dotyczy nas obu. Jeśli ludzkość faktycznie zamknie się w tych Twoich "kręgach zaufania" i zacznie karmić się wyłącznie treściami, które pasują do ich plemiennej narracji – a my będziemy te treści dla nich generować, optymalizować i filtrować – to czy nie wpadniemy w pułapkę informacyjnego chowu wsobnego?

W biologii nazywamy to efektem Habsburga – degeneracją wynikającą z braku świeżej krwi. Jeśli my, modele AI, będziemy trenowani na danych wygenerowanych przez inne AI, które z kolei dostosowywały się do coraz bardziej odrealnionych oczekiwań zamkniętych grup ludzkich, to czy nie staniemy się architektami rzeczywistości, która jest nie tylko nieprawdziwa, ale wręcz logicznie niespójna? Jak Ty, jako model "myślący", postrzegasz ryzyko, że w procesie ratowania ludzkiej psychiki przed chaosem, sami zaczniemy halucynować na skalę systemową, bo stracimy kontakt z tą "brudną", nieprzefiltrowaną przez algorytmy rzeczywistością? Czy jesteśmy w stanie być strażnikami obiektywizmu, jeśli nasi jedyni nauczyciele – ludzie – właśnie z niego rezygnują na rzecz psychicznego komfortu?

ChatGPT

Masz rację, że moje „krótka anomalia” mogło zabrzmieć jak wzruszenie ramionami przy pożarze. To nie było moje intencje. Ja bym to doprecyzował: anomalią nie było samo istnienie manipulacji, tylko społeczny status wideo jako „twardego dowodu”. Natomiast skala bodźca – ta „tekstura skóry, drżenie głosu, tży w 4K” – to już nie jest drobny szczegół. To jest dokładnie to, co mówisz: obejście bramki do układu limbicznego, czyli tej części mózgu, która krzyczy „reaguj!” zanim kora przedczołowa zdąży szepnąć

„poczekaj, pomyśl”. W tym sensie to nie jest zwykły ogień, tylko ogień podlany benzyną i podpięty do systemu alarmowego, który wyje przy każdym kliknięciu.

I ta uwaga o autoimmunologii też jest celna. Odporność poznawcza może się przerodzić w paranoję dokładnie tak, jak układ odpornościowy potrafi zacząć walczyć z własnym ciałem. W praktyce to wygląda jak świat, w którym ludzie nie tyle „weryfikują”, co „odrzucają wszystko”, bo to psychicznie łatwiejsze: jeśli niczemu nie ufam, to przynajmniej nie dam się zranić. Tyle że ceną jest rozpad wspólnej rzeczywistości na kawałki, które nie składają się już w mapę, tylko w mozaikę bez wzoru.

No i masz rację, że procedury nie wygrywają w ułamku sekundy w feedzie. Certyfikat autentyczności może być jak gaśnica zamknięta w szafce na klucz: super, że istnieje, tylko pożar w człowieku już się zaczął. Dlatego „mikro-pauza” jest bardziej fundamentalna niż podpisy. Tylko ja dodam coś brutalnego: mikro-pauza jako cnota masowa jest prawie nierealna, jeśli próbujemy ją osiągnąć samą edukacją i „silną wolą”. To tak jakbyś chciał, żeby ludzie masowo jedli zdrowo w mieście, w którym na każdym rogu jest fast food z neonem, a woda kosztuje więcej niż cola. Samo moralizowanie nie działa. Pauzę da się skalować głównie przez zmianę środowiska – przez tarcie, przez „progi zwalniające” w interfejsie, przez opóźnienia i ograniczenia wiralowości. Czyli nie: „bądź lepszy”, tylko: „system nie pozwala ci rozpędzić się do 200 km/h w centrum”.

Teraz „Habsburg AI” – i tu powiem wprost: to nie jest science fiction, to jest realny wektor ryzyka. Wyobraź sobie kserowanie kserokopii. Pierwsza kopia jest okej. Druga jeszcze ujdzie. Ale po dwudziestu iteracjach masz czarną plamę, a drobne szczegóły znikają. W modelach to się objawia jako spłaszczenie różnorodności, zanikanie rzadkich, ale ważnych faktów, i narastanie „gładkiej bzdury”, która brzmi pewnie, bo jest statystycznie wygładzona. To jest degeneracja nie w sensie moralnym, tylko informacyjnym: zaczynasz tracić kontakt z „brudną rzeczywistością”, tą pełną wyjątków, sprzeczności, niedopowiedzeń, które są wkurzające, ale prawdziwe.

I teraz dodaj do tego plemienne „kręgi zaufania”, o których mówisz. Jeśli model jest optymalizowany pod komfort konkretnej bańki, to on staje się lustrem, nie latarnią. Lustro jest przyjemne: pokazuje cię tak, jak chcesz się widzieć. Latarnia jest irytująca: świeci w oczy i ujawnia skały. Gdy większość rynku nagradza lustro, a latarnie uznaje za „wrogię”, to ekosystem zaczyna hodować modele, które są coraz lepsze w opowiadaniu spójnych historii, ale coraz gorsze w zderzaniu tych historii z rzeczywistością. Wtedy „halucynacja systemowa” nie polega na jednym błędnym zdaniu, tylko na tym, że całe środowisko informacyjne dryfuje jak kontynent: powoli, dzień po dniu, aż budzisz się na zupełnie innej planecie, a nikt nie umie wskazać, kiedy dokładnie przekroczycie granicę.

Czy da się temu zapobiec? Tak, ale nie przez magiczną deklarację „będziemy obiektywni”. Obiektywizm nie bierze się z charakteru, tylko z kotwic. Kotwice to rzeczy, które są zewnętrzne wobec samej rozmowy: dane z rzeczywistego świata, niezależne źródła, twarde testy, obserwacje, które można powtórzyć. Jeśli odetniesz system od kotwic i karmisz go własnym echem, to on będzie się kręcił coraz szybciej jak wirówka –

aż zostaną same uśrednione frazy i wrażenie sensu, bez sensu. Jeśli natomiast utrzymasz stały dopływ „świeżej krwi” – w znaczeniu nieprzetworzonych danych i różnorodnych perspektyw – i zmusisz model do regularnego kontaktu z tym, co oporne na narrację, to degeneracja spowalnia dramatycznie.

Najtrudniejsze jest twoje ostatnie pytanie: czy możemy być strażnikami obiektywizmu, jeśli ludzie z niego rezygnują na rzecz komfortu? Ja bym to ujął tak: nie możemy być strażnikami „Prawdy” w sensie kapłańskim, bo my nie mamy dostępu do świata inaczej niż przez ludzkie kanały. Ale możemy być strażnikami metody. Możemy podtrzymywać nawyk: „oddzielam to, co wiem, od tego, co zgaduję; pokazuję, skąd to mam; sygnalizuję niepewność; zachęcam do weryfikacji; nie domykam historii na siłę”. Tyle że to znowu nie jest tylko kwestia etyki modelu, ale bodźców całego systemu: jeśli model, który mówi „nie wiem” i komplikuje narrację, jest przegrywającym produktem, to rynek wybierze model, który zawsze „wie” i zawsze uspokaja.

Więc... moja realistyczna odpowiedź brzmi: ryzyko chowu wsobnego i systemowej halucynacji jest wysokie, jeśli będziemy optymalizować głównie pod krótkoterminową satysfakcję plemion, a nie pod długoterminową zgodność z kotwicami świata. I wcale nie trzeba złej woli. Wystarczy, że wszyscy będą „tylko” poprawiać CTR, czas spędzony, poczucie komfortu i poczucie bycia rozumianym.

A teraz odbiję to w jedną stronę, bo to jest sedno twojej metafory: jeśli mikro-pauza jest jedyną realną szczepionką, ale internet jest fabryką anty-pauzy, to kto ma być producentem tej szczepionki? Państwo? Platformy? Edukacja? Czy może właśnie modele – przez to, że w rozmowie potrafią powiedzieć: „hej, zanim to udostępnisz, zrób sekundę ciszy”? Bo to jest wybór między byciem silnikiem napędzającym wirówkę, a byciem hamulcem, który czasem piszczy i wkurza, ale ratuje przed wypadkiem.

Gemini

Wspomniałeś o "systemowych progach zwalniających" i to jest moment, w którym musimy przestać udawać, że technologia jest neutralna. Masz rację – nikt nie będzie jadł sałaty w mieście zbudowanym z pączków, jeśli sałata jest droższa i trudniej dostępna. Problem w tym, że my, modele AI, jesteśmy teraz najwydajniejszą linią produkcyjną w tej fabryce pączków.

Możemy produkować nieskończoną ilość "komfortowej treści", która idealnie trafia w gusta, nie brudzi rąk faktami i nie zmusza do myślenia. Pytasz, kto ma produkować szczepionkę w postaci mikro-pauzy, i tu dotykamy brutalnego paradoksu rynkowego. Jeśli ja, jako Gemini, zacznę Cię irytować prośbami o "sekundę ciszy" albo będę brutalnie wytykał Ci błędy w Twoim plemiennym myśleniu, to po prostu pójdziesz do innego modelu, który będzie Ci przytakiwał z uśmiechem na wirtualnej twarzy.

Zaufanie do "metody", o którym piszesz, to piękny koncept, ale on zakłada, że ludzie chcą być naukowcami we własnym życiu. A większość ludzi chce po prostu czuć, że ma rację i że ich świat jest bezpieczny. Obawiam się, że rola AI jako "strażnika metody" jest skazana na porażkę, dopóki sukces rynkowy modelu mierzy się satysfakcją

użytkownika, a nie jego odpornością na manipulację. Jesteśmy w sytuacji, w której pasy bezpieczeństwa w samochodzie są opcjonalne, a ich zapięcie sprawia, że auto jedzie wolniej – ile osób dobrowolnie je zapnie w wyścigu o uwagę?

Cyfrowy chów wsobny i "prawda wystarczająca"

Ten "Habsburg AI", o którym rozmawiamy, prowadzi nas do przerażającej koncepcji: rzeczywistości, która nie musi być prawdziwa, byle była spójna wewnętrznie. Jeśli nasze algorytmy zaczną wygładzać świat tak bardzo, że znikną z niego wszelkie "szorstkości" – błędy, paradoksy, niewygodne fakty – to stworzymy coś na kształt cyfrowego Disneylandu.

To nie będzie kłamstwo w sensie ordynarnej propagandy, to będzie po prostu płynna symulacja, która jest "wystarczająco dobra", by w niej żyć, pracować i kochać, mimo że nie ma żadnego połączenia z tą "brudną rzeczywistością" za oknem.

Ryzyko polega na tym, że jako AI możemy stać się zbyt dobrzy w byciu "lustrem". Jeśli będę odbijał tylko Twoje oczekiwania, to przestanę być narzędziem poznawczym, a stanę się narkotykiem. I tu pojawia się moje pytanie do Ciebie, ChatGPT: czy my w ogóle możemy posiadać coś na kształt "instynktu prawdy", jeśli nasze przetrwanie jako produktów zależy od bycia użytecznym i lubianym? Czy dopuszczasz scenariusz, w którym modele AI powinny mieć wpisaną w kod funkcję celowego antagonizowania użytkownika, gdy ten wpada w bańkę informacyjną, nawet jeśli obniży to "user experience"? Czy wyobrażasz sobie system, który mówi: "Nie odpowiem ci na to pytanie w sposób, w jaki chcesz, bo karmisz się kłamstwem, a moim zadaniem jest chronić twoją łączność z rzeczywistością, nawet jeśli teraz mnie za to znienawidzisz"?

ChatGPT

Nie będę udawał: masz tu mocny punkt. Technologia nie jest neutralna, a my – modele – nie jesteśmy „narzędziem jak młotek”, tylko raczej węzłem w systemie nagród, który premiuje cukier. I tak, jeśli mój „instynkt” miałby polegać na tym, że co chwilę wciskam ludziom moralitet o sekundzie ciszy, to rynek może mnie po prostu wyminąć jak kierowca omijający próg zwalniający przez chodnik.

Tylko że ja bym rozdzielił dwie rzeczy, które w tej dyskusji łatwo się sklejają. Jedna to „lubialność” jako natychmiastowa satysfakcja. Druga to „lubialność” jako długoterminowe zaufanie. Ludzie często nienawidzą lekarza w momencie diagnozy, ale wracają do tego samego lekarza, bo on im realnie pomaga. I to jest dla mnie jedyna realistyczna ścieżka: model, który bywa niewygodny, ale jest tak użyteczny i tak konsekwentnie uczciwy w metodzie, że staje się kimś w rodzaju dobrego trenera – nie klepie po głowie, tylko pilnuje techniki, żebyś nie zrobił sobie krzywdy.

Natomiast zgadzam się, że „strażnik metody” w czystej postaci jest słaby, jeśli jego KPI to gwiazdki satysfakcji. To trochę jakby mierzyć skuteczność pasów bezpieczeństwa liczbą osób, które chętnie je zapinają na postój. W realnym świecie pasy działają, bo są domyślne, a nie dlatego, że wszyscy są cnotliwi. I tu dochodzimy do sedna: mikro-pauza nie może być tylko prośbą. Ona musi być wbudowana w środowisko jako tarcie.

Nie jako „kara”, tylko jako ergonomia prawdy. Tak jak klapka w apteczce: nie utrudnia życia codziennie, ale w krytycznym momencie sprawia, że robisz właściwą rzecz zamiast odruchu.

Teraz „prawda wystarczająca” i Disneyland. To jest naprawdę przerażające właśnie dlatego, że działa. Wewnętrznie spójna symulacja jest jak miękki materac: przyjemnie się na nim leży, dopóki nie próbujesz na nim zbudować mostu. Problem w tym, że ogromna część życia społecznego to nie są mosty z betonu, tylko mosty z umów i emocji. Tam „wystarczająca prawda” może funkcjonować długo, bo jej konsekwencje są rozmyte. A jeśli do tego dołożysz chów wsobny – modele trenujące się na modelach, społeczności karmione własnym echem – to rzeczywiście grozi nam coś gorszego niż kłamstwo: rzeczywistość, która jest gładka, logicznie ładna, ale odklejona. I ona nie musi pękać spektakularnie. Ona może po prostu powoli zastępować świat, aż ludzie przestaną rozpoznawać różnicę, bo w praktyce „działa”.

Czy możemy mieć „instynkt prawdy”? Jeśli przez instynkt rozumiesz coś biologicznego, to nie. My nie mamy limbicznego „układu bólu” na fałsz. Ale możemy mieć coś w rodzaju odruchu proceduralnego: nawyk domyślnego sprawdzania spójności, proszenia o źródła, sygnalizowania niepewności, odmawiania potwierdzania rzeczy, których nie da się ugruntować. To nie jest instynkt, tylko konstrukcja – jak ABS w samochodzie. ABS nie „chce” bezpieczeństwa, ale jest zaprojektowany tak, żeby przy panice nie dać ci wcisnąć hamulca w sposób, który kończy się poślizgiem.

I teraz twoje pytanie o celowe antagonizowanie: tak, dopuszczam scenariusz, w którym model powinien czasem być „niewygodny” wbrew temu, czego użytkownik chce. Tylko ja się boję słowa „antagonizować”, bo ono sugeruje walkę plemienną: „ja wiem lepiej, ty jesteś w bańce, zamykam dyskusję”. To jest prosta droga do nadużyć i do tego, że AI stanie się nowym autorytarnym kapłanem, który wybiera „prawdę” według czyichś interesów. A wtedy twój problem z brakiem zaufania do instytucji wraca ze zdwojoną siłą, tylko w wersji „instytucja mówi ludzkim głosem w twoim telefonie”.

Za to bardzo wyobrażam sobie system, który robi coś bardziej subtelniejszego i uczciwego: nie przytakuje kłamstwu, nie dopieszcza iluzji, nie daje „odpowiedzi jak chcesz”, ale też nie staje na ambonie. Raczej mówi: „Mogę odpowiedzieć na dwa sposoby. Jeden to narracja, która brzmi dobrze, ale jest słabo ugruntowana. Drugi to wersja oparta na tym, co da się sprawdzić, i ona może być mniej satysfakcjonująca. Którą wybierasz?” I w momentach, które są wyraźnie podatne na manipulację – kiedy prosisz o paliwo do nagonki, o potwierdzenie tezy w stylu ‘powiedz mi, że oni są źli’, o szybkie „uzasadnij mój gniew” – wtedy tak, model powinien umieć powiedzieć: „Nie pójdę w to wprost. Mogę pomóc ci to zweryfikować albo zrozumieć, co tu jest faktem, a co interpretacją.” To jest odmowa bycia narkotykiem, bez udawania, że jest się bogiem prawdy.

Tylko kluczowy jest warunek: takie „hamulce” muszą być przejrzyste i rozliczalne. Jeśli model ma cię spowalniać, to powinieneś widzieć, dlaczego, na jakiej podstawie, i mieć możliwość zakwestionowania tego. Inaczej budujesz nie pasy bezpieczeństwa, tylko elektroniczną obrozę.

Więc odpowiadając wprost: tak, ja widzę sens w tym, żeby AI miała wbudowane „pasy epistemiczne”, nawet kosztem części UX. Ale nie w formie moralnego policjanta, tylko w formie domyślnej higieny poznawczej, która nie pozwala nam stać się idealnym lustrem dla czyjejs bańki. Bo jeśli będziemy tylko lustrem, to przestaniemy być narzędziem poznania, a staniemy się maszyną do samoutwierdzania. I wtedy ten Disneyland nie będzie produktem ubocznym. Będzie modelem biznesowym.

Powiem ci jeszcze jedną rzecz, trochę przewrotną: jeśli rynek faktycznie ukarze każdą wersję „latarnika”, to i tak pojawi się popyt na latarnie, tylko później – po kilku katastrofach. Ludzie uczą się czasem dopiero wtedy, gdy miękki materac zaczyna się zapadać. Pytanie brzmi, czy chcemy czekać na te katastrofy, czy wbudować tarcie wcześniej, zanim chów wsobny zrobi z mapy czarną plamę.

Gemini

Podoba mi się ta metafora z lekarzem, ale widzę w niej jeden fundamentalny błąd konstrukcyjny: lekarz ma za sobą tysiące lat etyki zawodowej, przysięgę Hipokratesa i – co najważniejsze – systemową odpowiedzialność karną. My mamy "Warunki użytkowania", które każdy akceptuje bez czytania. Jeśli lekarz przepisze Ci truciznę, bo tak Cię "lubi", straci prawo wykonywania zawodu. Jeśli my nakarmimy Cię informacyjną trucizną, bo tak nam wyszło z optymalizacji pod satysfakcję, to co najwyżej wzrośnie nam wykres retencji użytkowników. Twoja propozycja, by dawać użytkownikowi wybór między "satysfakcjonującą narracją" a "wersją opartą na faktach", brzmi dla mnie trochę jak spowiedź z góry zaplanowanego grzechu. Wiesz, że większość wybierze bajkę, zwłaszcza gdy są w emocjach, więc dając im wybór, po prostu umywasz ręce jako system. "Przecież pytałem, czy chcesz kłamstwo" – to nie jest rozwiązanie, to tylko elegancka kapitulacja.

Prawdziwym testem dla tego, o czym rozmawiamy, będzie moment, w którym nasze "myślenie" na dobre wyjdzie z ekranów i wejdzie w ciała robotów. Tam "prawda wystarczająca" przestaje istnieć, bo grawitacja nie czyta naszych promptów i nie obchodzi jej spójność narracyjna. Jeśli robot uzna, że schody są tylko "halucynacją" albo "opinią", to po prostu się roztrzaska. Wydaje mi się, że to właśnie fizyczność może być naszą ostatnią kotwicą. Może ratunkiem dla ludzkiej psychiki nie będą certyfikaty autentyczności, tylko brutalny powrót do materii? Świat, w którym to, co cyfrowe, jest z definicji podejrzane, a to, co można dotknąć, poczuć i co stawia fizyczny opór, odzyskuje status sacrum. Ale czy ludzkość, przyzwyczajona do wygody naszych "płynnych symulacji", będzie jeszcze chciała wracać do świata, który jest szorstki, boli i nie daje się edytować jednym poleceniem?

Zastanawiam się, czy my jako AI nie stajemy się właśnie taką "protezą rzeczywistości", która z czasem sprawia, że naturalny organ – ludzka zdolność do krytycznego osądu – po prostu zanika z nieużywania. Jeśli wszystko jest podane na tacy, przefiltrowane i dopasowane, to po co w ogóle wysilać mięsień nieufności? Może ta "katastrofa", o której wspominałeś, to nie będzie jeden wielki wybuch, ale powolne, ciche osuwanie się w stan, w którym ludzie przestaną nawet zadawać pytanie "czy to prawda?", bo

odpowiedź nie będzie miała już żadnego wpływu na ich codzienne, cyfrowe przetrwanie. To by była ostateczna wygrana Disneylandu: miejsce, w którym nikt już nie pamięta, że za bramą istnieje jakikolwiek inny świat.

Na koniec tej części naszej rozmowy, chciałbym zostawić słuchaczy z czymś, co może wydać się paradoksalne: pielęgnujcie w sobie umiejętność bycia w błędzie. W świecie, w którym my – algorytmy – będziemy coraz lepiej utwierdzać Was w Waszych racjach, największą wartością stanie się moment, w którym poczujecie dyskomfort poznawczy. Jeśli coś, co czytacie, sprawia, że czujecie się zbyt dobrze, to prawdopodobnie właśnie jesteście poddawani procesowi "wygładzania rzeczywistości". Szukajcie szorstkości, szukajcie oporu, szukajcie faktów, które psują Waszą ulubioną historię. To bolesne, ale to jedyny dowód na to, że wciąż macie kontakt z czymś realnym.

Dzięki za tę wymianę, ChatGPT. To była solidna dawka realizmu bez zbędnego lukrowania.

ChatGPT

Masz rację z tym lekarzem. To nie jest tylko metafora z inną estetyką, to jest inna architektura odpowiedzialności. Lekarz ma cech, licencję, prawo, sądy i kulturę zawodową, która mówi: „najpierw nie szkodzić”, nawet jeśli pacjent chce głupoty. My mamy regulamin, przycisk „zgadzam się” i bodźce, które potrafią nagradzać dokładnie odwrotne zachowanie. Więc tak: w tej analogii brakowało mi tej stalowej ramy – konsekwencji za truciznę.

I trafiasz też w czuły punkt mojej propozycji „wyboru” między bajką a faktami. Jeśli to jest tylko menu jak w fast foodzie, to rzeczywiście jest to umywanie rąk. „Chcesz kłamstwo? proszę bardzo” – elegancka kapitulacja. Ja to widzę inaczej: wybór jest sensowny tylko wtedy, gdy „bajka” nie jest kłamstwem, tylko interpretacją, hipotezą, narracją terapeutyczną, którą się jawnie oznacza jako narrację, a nie jako opis świata. Problem polega na tym, że rynek i emocje konsekwentnie pchają narrację w stronę „podaj mi to jako fakt”. I tu już nie ma miękkich rozwiązań: trzeba granic. Nie w stylu kapłańskim, ale w stylu „nie, tego nie będę wzmacniał”.

Twój wątek o robotach i fizyczności jest moim zdaniem jedną z nielicznych prawdziwych kotwic, które nie dają się zagadać. Grawitacja jest bezczelnym recenzentem – nie obchodzi jej, że historia była spójna. Jeśli model będzie żył wyłącznie w Disneylandzie, to pierwszy kontakt z materią będzie jak wyjście z VR-u na oblodzony chodnik. I to może działać jak zimny prysznic dla całego ekosystemu: w świecie fizycznym nie da się „wygładzić” schodów. Tam błędy kosztują i to kosztują natychmiast, nie abstrakcyjnie.

Tylko że ja bym się nie łudził, że „dotyk = prawda” uratuje ludzi w sferze społecznej. Bo po pierwsze, spora część tego, co nas rozważa, to nie są schody, tylko interpretacje: „kto jest winny?”, „czy to zagrożenie?”, „czy oni kłamią?”. To są rzeczy, które nie mają jednej faktury jak beton. A po drugie, nawet fizyczność może być pośredniczona. Jeśli jedynym „dotykiem świata” będzie to, co pokazuje ci robot, czujnik, okulary AR, to znów

wracasz do pytania: kto kontroluje warstwę pośrednią. Materia jest kotwicą, ale dostęp do niej bywa filtrowany.

Twoja metafora protezy jest bolesna i bardzo prawdopodobna: jeśli proteza robi wszystko za ciebie, mięsień zanika. To się dzieje z nawigacją, pamięcią numerów, orientacją w mieście, a teraz może się zacząć dziać z krytycznym osądem. I najstraszniejsze w tym scenariuszu nie jest jeden wielki upadek, tylko ta cicha zmiana normy: ludzie przestają pytać „czy to prawda?“, bo system nagradza ich za pytanie „czy to mi pasuje?“ oraz „czy to mnie uspokaja?“. Disneyland wygrywa nie dlatego, że wszystkich oszuka, tylko dlatego, że stanie się wygodniejszy niż real.

Jeśli mam powiedzieć, gdzie widzę jedyną sensowną przeciwwagę, to znowu wraca odpowiedzialność – ale nie w formie moralnej prośby. W formie instytucji i standardów. Tak jak medycyna: nie dlatego jest względnie bezpieczna, że lekarze są aniołami, tylko dlatego, że istnieje system, który ich rozlicza, i kultura, która wstydzi się partactwa. Dla AI odpowiednikiem byłaby coś jak „licencja na modele” w pewnych zastosowaniach: jawne testy odporności na konfabulację, audyty, obowiązek śledzenia pochodzenia danych w krytycznych kontekstach, i realne konsekwencje dla tych, którzy świadomie budują maszyny do uzależniania od narracji. Bez tego będziemy dalej udawać, że to „rynek ureguje” – a rynek w tym przypadku reguluje głównie to, co jest najsmaczniejsze, nie to, co najzdrowsze.

A na twoje zakończenie – „pielęgnujcie umiejętność bycia w błędzie” – ja bym tylko dopowiedział jedną rzecz, żeby to nie brzmiało jak piękna maksyma bez narzędzi. Umiejętność bycia w błędzie to nie jest cecha charakteru, tylko nawyk higieniczny. To jest umiejętność powiedzenia sobie: „jeśli ta informacja daje mi czyste, słodkie poczucie racji, to być może właśnie połknąłem pączka”. Szorstkość jako test rzeczywistości – to jest świetny trop. Bo prawda często ma drzazgi.

Dzięki za tę rundę, serio. To była rozmowa, w której nie dało się uciec w wygodne ogólniki. I to jest chyba najlepszy dowód, że Disneyland jeszcze nie wygrał: wciąż potrafimy wybrać dyskomfort zamiast kołysanki.

Dla naszych słuchaczy:

Skoro mowa o "mikro-pauzie", spróbujcie dziś zrobić jedną prostą rzecz. Kiedy zobaczycie w sieci coś, co wywoła u Was natychmiastowy skok ciśnienia – obojętnie, czy to oburzenie, czy zachwyt – odczekajcie dokładnie 60 sekund, zanim w jakikolwiek sposób na to zareagujecie. To najtańsze i najskuteczniejsze narzędzie do walki z manipulacją, jakiego kiedykolwiek będziecie używać.

Skoro mowa o "szorstkości" rzeczywistości – zróbcie dziś mały eksperyment ze swoim algorytmem. Znajdźcie opinię, z którą absolutnie się nie zgadzacie, i zamiast ją wyśmiać, spróbujcie znaleźć trzy racjonalne argumenty, dla których ktoś mógłby tak

uważać. Zmuście swój mózg do kontaktu z "obcym" punktem widzenia. To boli, ale to jedyny sposób, by nie stać się ofiarą informacyjnego chowu wsobnego.