Era Agentów: Od Narzędzi do Cyfrowych Pracowników

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje fundamentalną zmianę w postrzeganiu sztucznej inteligencji – przejście od prostych narzędzi sterowanych promptami do autonomicznych agentów pełniących rolę cyfrowych pracowników. Autor szczegółowo omawia techniczne aspekty tej transformacji, w tym architekturę RAG, wykorzystanie embeddingów oraz znaczenie Model Context Protocol (MCP) w budowaniu bezpiecznych i otwartych ekosystemów AI. Tekst podkreśla, że AI przestaje być jedynie dodatkiem, a staje się nową warstwą organizacji poznawczej, przejmującą funkcje predykcyjne i decyzyjne. Wskazuje na konieczność odejścia od magicznego postrzegania promptów na rzecz precyzyjnej inżynierii systemowej. Omówione zostają również wyzwania związane z bezpieczeństwem, jurysdykcją danych oraz unikaniem zjawiska vendor lock-in w nowoczesnej architekturze semantycznej. To kluczowa lektura dla osób projektujących przyszłość interakcji człowiek-maszyna.

This article examines the fundamental shift in the perception of artificial intelligence—the shift from simple prompt-driven tools to autonomous agents acting as digital workers. The author provides a detailed discussion of the technical aspects of this transformation, including the RAG architecture, the use of embeddings, and the importance of the Model Context Protocol (MCP) in building secure and open AI ecosystems. The article emphasizes that AI is no longer a mere add-on but a new layer of cognitive organization, taking on predictive and decision-making functions. It highlights the need to shift from a magical perception of prompts to precise systems engineering. The challenges of security, data jurisdiction, and avoiding vendor lock-in in modern semantic architecture are also discussed. This is essential reading for those designing the future of human-machine interaction.

Współczesna debata o sztucznej inteligencji przesuwa się z fascynacji magią językowych modeli ku surowej inżynierii systemowej. Artykuł dowodzi, że AI nie jest już jedynie

przezroczystym narzędziem do generowania treści, lecz staje się autonomicznym aparatem kompozycji decyzji, który wymusza na nas stworzenie nowych ram operacyjnych. Autorzy przekonują, że kluczowe technologie – takie jak RAG, embeddingi czy protokół MCP – przestają być technicznymi ciekawostkami, a stają się fundamentem nowej ontologii sprawczości. W tym modelu prompt przestaje być „życzeniem" do cyfrowej studni, a staje się konstytucją zadania, która musi być spójna z rygorami biznesowymi, prawnymi i bezpieczeństwa. Główna teza tekstu wskazuje, że przyszłość aplikacji AI nie należy do twórców najbardziej elokwentnych czatów, lecz do architektów potrafiących zaprojektować bezpieczne granice dla cyfrowej autonomii. W dobie AI Act i rosnącej złożoności systemów, sukces zależy od przejścia od technokratycznego entuzjazmu ku dojrzałej administracji procesów. Technologia traci swój magiczny status, stając się elementem ustrojowym, który wymaga od nas nie tyle zachwytu, co precyzyjnego zarządzania ryzykiem i odpowiedzialnością.

SŁOWA KLUCZOWE

Era Agentów | cyfrowy pracownik | inżynieria promptów | Model Context Protocol | RAG | embeddingi | przestrzeń wektorowa | miara cosinusowa | vendor lock-in | walidacja danych | backend | multimodalność | chunking | architektura semantyczna | sprawczość systemów autonomicznych

Od magii promptów do inżynierii cyfrowych pracowników

Każda epoka ma swoją naiwną metaforę technologiczną, która służy za intelektualną protezę w obliczu nadchodzącej zmiany. Wczoraj była nią strona internetowa pojmowana jako statyczna witryna, potem aplikacja rozumiana jako poręczne narzędzie, dziś zaś coraz częściej bywa nią agent, czyli cyfrowy pracownik zdolny do podejmowania decyzji i wykonywania zadań w imieniu użytkownika. Zmiana ta nie ma charakteru kosmetycznego, gdyż dotyka samej ontologii systemu, czyli nauki o strukturze bytu i relacjach wewnątrz oprogramowania. Nie budujemy już wyłącznie narzędzi dla użytkownika, lecz projektujemy całe środowiska operacyjne, w których sztuczna inteligencja pracuje pod jego mandatem. W tym nowym układzie AI działa w granicach interesu

człowieka i z użyciem zasobów, do których uzyskała jedynie warunkowy dostęp. Przyszłość aplikacji nie jest zatem historią o coraz ładniejszym czacie, lecz opowieścią o delegowaniu sprawczości na systemy autonomiczne.

Właśnie dlatego architektura, bezpieczeństwo, jurysdykcja danych oraz ekonomia integracji przestają być nudnym zapleczem, a stają się absolutnym centrum zagadnienia. W oficjalnych materiałach OpenAI dotyczących narzędzi i Responses API widać już wyraźnie, że model ma nie tylko generować tekst, lecz także aktywnie korzystać z wyszukiwania, plików i zdalnych serwerów. Równolegle Anthropic opisał MCP, czyli Model Context Protocol, będący otwartym standardem bezpiecznych połączeń między modelami a zewnętrznymi źródłami danych, jako fundament nowej ekologii cyfrowej. To nie jest jedynie techniczny margines rynku, lecz fundamentalna zmiana jego języka i sposobu operowania informacją. Rozwiązuje to problem chaosu integracyjnego i zapobiega uzależnieniu od jednego dostawcy, co w świecie inżynierii nazywamy zjawiskiem vendor lock-in. Wszak przejście od prostych odpowiedzi do realnego działania wymaga solidnych, ustandaryzowanych mostów między kodem a rzeczywistością.

Na tym tle trafna okazuje się intuicja obecna w materiałach Fundacji Dobre Państwo, sugerująca, że AI nie jest jedynie dodatkiem do istniejących instytucji poznawczych. Sztuczna inteligencja staje się nową warstwą organizacji przewidywania, selekcji i rozdziału sprawczości w nowoczesnym społeczeństwie. W tekstach o agencie racjonalnym podkreślono, że racjonalność w sensie technicznym przestaje być metaforą i staje się konkretnym zadaniem inżynieryjnym. Z kolei w rozważaniach o mechanice postępu postawiono tezę, że technologia sama w sobie nie posiada teleologii, czyli wewnętrznego, zdefiniowanego celu. Może ona prowadzić do stagnacji, jeśli nasze instytucje będą nagradzać wyłącznie eksploatację tego, co już dobrze znane i bezpieczne. Jeszcze mocniej wybrzmiewa to w materiale o życiu jako obliczeniu, gdzie AI zostaje opisana jako nowa warstwa organizacji poznawczej, przejmująca funkcje predykcyjne.

To ważne, bo przesuwa nas od banału w rodzaju „model odpowiada na pytania” ku znacznie poważniejszej i bardziej wymagającej tezie. System AI nie jest dziś tylko maszyną tekstową, lecz aparatem kompozycji decyzji, który spina epistemologię z ekonomią kosztów transakcyjnych. Współczesne systemy muszą godzić prawo dostępu z kulturowymi modelami delegowania odpowiedzialności, co czyni je strukturami niezwykle złożonymi. AI staje się nową warstwą organizacji poznawczej, która nie tylko dostarcza informacji, ale współkształtuje architekturę zaufania społecznego. W tym sensie technologia przestaje być przezroczystym narzędziem, a staje się aktywnym uczestnikiem procesów decyzyjnych. Reszta jest marketingiem, który próbuje ubrać tę surową rzeczywistość w piórka przystępnych, kolorowych interfejsów.

Warto więc rozpocząć od pojęcia promptu, ponieważ jest ono w debacie publicznej notorycznie i szkodliwie banalizowane. W języku potocznym prompt bywa traktowany jak życzenie wrzucone do cyfrowej studni, co rodzi fałszywe przekonanie o magicznej naturze tej technologii. Tymczasem w dojrzałej architekturze AI prompt nie jest kaprysem użytkownika, tylko precyzyjnym instrumentem zarządzania kontekstem. Określa on rolę modelu, zakres zadania, pożądaną strukturę odpowiedzi oraz dopuszczalne źródła, z których system może czerpać wiedzę. Jest zarazem instrukcją semantyczną i mechanizmem kontraktowym, który definiuje granice bezpieczeństwa oraz format wyjścia. Dobry prompt nie tyle „prosi" o wynik, ile ustanawia twarde warunki wykonania powierzonego zadania.

Z tego powodu współczesna inżynieria promptów nie polega już na szukaniu magicznych zaklęć w mediach społecznościowych, lecz na surowej dyscyplinie projektowej. Frameworki i biblioteki programistyczne eksponują funkcje generowania tekstu oraz wywoływania narzędzi, ponieważ prompt musi współpracować z całym potokiem wykonawczym aplikacji. Dokumentacja Vercel AI SDK wprost opisuje funkcje takie jak generateText jako rdzeń generowania odpowiedzi, będący właściwym narzędziem dla agentów korzystających z zewnętrznych zasobów. Prompt staje się konstytucją zadania, która musi być spójna z logiką biznesową i technicznymi ograniczeniami systemu. Nie ma tu miejsca na poezję, jest za to mnóstwo miejsca na precyzyjną strukturę danych. Postęp bywa komiczny, zwłaszcza gdy uświadomimy sobie, jak bardzo technologia wymusza na nas powrót do ścisłego definiowania pojęć.

Tu pojawia się pierwszy moment humoru, w dodatku humoru dość okrutnego, który obnaża nasze przywiązanie do myślenia magicznego. W wielu organizacjach prompt traktuje się jak nowoczesny odpowiednik średniowiecznej formuły „niech się stanie", oczekując cudów od kilku niedbałych słów. Ktoś wpisuje do modelu polecenie: „zrób strategię, ale dokładnie i profesjonalnie", po czym z miną audytora z Olimpu oczekuje prawdy, trafności i świeżo parzonej kawy. Model w takiej sytuacji przypomina urzędnika wszechświata, któremu kazano przygotować plan reformy podatkowej na poplamionej serwetce z kawiarni. Biedny algorytm nie ma danych, nie posiada dostępu do systemów ani uprawnień, lecz za to mierzy się z niebotycznymi oczekiwaniami co do tonu wypowiedzi. Jeśli coś tutaj jest absurdalne, to nie model, lecz przekonanie, że nieprecyzyjny rozkaz może zrodzić precyzyjny rezultat.

W nauce byłoby to śmieszne. W biznesie bywa drogie. W prawie kończy się sporem, który rzadko przynosi satysfakcję którejkolwiek ze stron. W systemach agentowych takie podejście może skończyć się błędnym działaniem wykonanym z pełną, proceduralną grzecznością i nienaganną składnią. Cywilizacja zbudowała lokaja, a potem musiała mu tłumaczyć, że nie każda karteczka wsunięta pod drzwi jest wiążącą instrukcją od właściciela domu. Musimy zrozumieć, że inteligencja

bez struktury i danych jest jedynie elokwentną pustką, która potrafi pięknie błądzić. Dlatego właśnie inżynieria promptów musi ewoluować w stronę projektowania systemowego, gdzie każde słowo ma swoją wagę i funkcję w architekturze. Bez tego pozostaniemy w sferze cyfrowego wróżbiarstwa, które z prawdziwą efektywnością ma niewiele wspólnego.

Dlatego nowoczesna aplikacja AI nie może opierać się na samym promptingu, choćby był on najbardziej wyrafinowany i błyskotliwy. Potrzebuje ona solidnego API, pamięci roboczej, walidacji oraz rozbudowanej warstwy narzędziowej, która osadzi ją w rzeczywistości. API pozostaje absolutnym fundamentem tej architektury, ponieważ rozwiązuje elementarny problem rozdziału obowiązków między interfejsem a logiką biznesową. Współczesna praktyka inżynieryjna jest tu bezlitosna dla amatorstwa i nie znosi drogi na skróty w kwestiach bezpieczeństwa. Klucze uwierzytelniające nie mogą żyć po stronie klienta, a żądania do modeli nie powinny wychodzić bezpośrednio z przeglądarki użytkownika. Sekrety muszą być przechowywane w bezpiecznym środowisku serwerowym, a publiczne punkty końcowe wymagają rygorystycznej walidacji i obserwowalności.

Nie jest to kwestia estetyki kodu, lecz twardego rachunku ekonomicznego i odpowiedzialności prawnej za działanie systemu. OpenAI oraz Vercel przesuwają ciężar z prostego wysyłania promptów ku kontrolowanemu przepływowi przez backend i zaawansowane narzędzia. To backend decyduje, kiedy wolno użyć konkretnego modelu, jakie dane można bezpiecznie ujawnić i który mechanizm awaryjny uruchomić w razie błędu. W tym procesie kluczowe stają się embeddings, czyli technika geometryzacji podobieństwa semantycznego poprzez mapowanie danych do przestrzeni wektorowej. Pozwalają one systemowi „rozumieć" kontekst poprzez matematyczne relacje między pojęciami, a nie tylko przez proste dopasowanie słów kluczowych. Dzięki temu aplikacja może operować na znaczeniach, co jest fundamentem nowoczesnego wyszukiwania i wnioskowania.

W tym miejscu trzeba jasno powiedzieć rzecz dla wielu entuzjastów nieprzyjemną, a może nawet nieco bolesną. AI nie znosi architektury, ona ją dopiero brutalnie wymusza na każdym, kto chce wyjść poza etap zabawkowego czatu. Im bardziej system wydaje się „magiczny" i płynny w działaniu, tym bardziej wymaga przyziemnych, wręcz nudnych urządzeń kontroli. W klasycznym internecie błędnie zaprojektowany przycisk co najwyżej irytował użytkownika, ale w aplikacji agentowej błędna decyzja może mieć realne skutki. System może zamówić niechciany towar, skasować ważne dane lub ujawnić poufne informacje w cudzym imieniu, działając przy tym z pełnym przekonaniem o własnej racji. Stąd bierze się renesans starej mądrości inżynierskiej, która zawsze stawiała rygor ponad entuzjazm.

Prawdziwie nowoczesne nie jest to, co najgłośniej krzyczy o autonomii i cyfrowej wolności, ale to, co najdokładniej projektuje granice tej autonomii. W tym sensie najlepsze systemy agentowe będą przypominały nie rozkrzyczane startupowe prezentacje, lecz dobrze napisane konstytucje operacyjne. Musimy pogodzić się z faktem, że sprawczość maszyn wymaga od nas większej precyzji niż kiedykolwiek wcześniej. Każdy element systemu, od RAG, czyli architektury łączącej generowanie tekstu z dynamicznym wyszukiwaniem informacji, po mechanizmy walidacji, musi współgrać w ramach większej całości. Tylko wtedy będziemy mogli mówić o budowaniu zaufania do systemów, które mają za nas decydować i działać. Reszta to tylko szum informacyjny, który opadnie, gdy przyjdzie do rozliczania realnych efektów pracy cyfrowych agentów.

Embeddings: Architektura semantyczna i władza pojęciowa

Na tym tle szczególne znaczenie zyskują embeddingi, czyli osadzenia, które w popularnym dyskursie bywają redukowane do roli tajemniczych ciągów cyfr rzekomo „rozumiejących" treść. Każda epoka ma swoją naiwną metaforę technologiczną, niemniej tutaj lepiej mówić o geometryzacji podobieństwa semantycznego. Model matematyczny mapuje tekst, obrazy czy dźwięki do przestrzeni wektorowej, w której obiekty o zbliżonym znaczeniu lądują blisko siebie. Pozwala to operować na sensie, a nie tylko na znakach, co OpenAI wykorzystuje do klasyfikacji czy wykrywania anomalii. Do porównywania tych pozycji rekomendowana jest miara cosinusowa (cosine similarity), czyli miara kątowa określająca bliskość wektorów w przestrzeni. Prompt staje się konstytucją zadania, a embeddingi stają się geometrią sensu.

Rok 2026 przynosi rozszerzenie tego paradygmatu ku rozwiązaniom multimodalnym, co widać w modelu Gemini Embedding 2 od Google. System ten mapuje tekst, wideo oraz dokumenty do wspólnej przestrzeni reprezentacji, zacierając granice między formami zapisu danych. Architektura semantyczna przestaje być zatem sztuczką, stając się fundamentem infrastruktury poznawczej dla systemów pracujących na wielu modalnościach. Nie jest to już tylko narzędzie, lecz próba skatalogowania świata w sposób, który maszyna może przeliczyć. W środowisku naukowym budziłoby to jedynie pobłażliwy uśmiech, lecz w biznesie generuje realne koszty, a w prawie kończy się sporem. Ewolucja techniczna miewa w sobie coś z ironii, zwłaszcza gdy kultura zostaje sprowadzona do współrzędnych na wykresie.

Embeddingi nie stanowią wiedzy, lecz są warunkiem porządkowania dostępu do niej, co ma wagę filozoficzną. W sensie epistemologicznym, czyli dotyczącym teorii poznania, osadzenia te nie przechowują prawdy, a jedynie strukturę podobieństw. Ekonomicznie nie dają odpowiedzi, lecz

obniżają koszt odnalezienia kandydatów do jej udzielenia. Technologia ta nie znosi konfliktów interpretacyjnych, lecz przenosi je do warstwy tego, co algorytm uznał za „bliskie". Kto projektuje proces tworzenia tych osadzeń, ten projektuje nową ontologię podobieństwa, sprawując subtelną władzę pojęciową. Sztuczna inteligencja nie tyle unika sztywnej architektury, co bezlitośnie ją na nas wymusza, zamieniając dawną magię w surową politykę widzialności.

Architektura RAG i narodziny cyfrowej sprawczości

RAG, czyli retrieval-augmented generation, to termin, który zrobił w ostatnich latach oszałamiającą, choć momentami nieco podejrzaną karierę, obiecując nam jednocześnie świeżość wiedzy i radykalne ograniczenie halucynacji. Ta obietnica brzmi nadzwyczaj rozsądnie, niemniej staje się ona operacyjnie prawdziwa tylko wtedy, gdy z pokorą zrozumiemy jej techniczne oraz ontologiczne granice. Oficjalne materiały OpenAI opisują retrieval jako mechanizm oparty na bazach wektorowych (vector stores) i semantycznym wyszukiwaniu, przy czym narzędzia te mają regulować bilans między jakością a kosztami. Wszak RAG nie jest żadną cudowną protezą prawdy, lecz raczej wyrafinowaną architekturą zasilania modelu relewantnym kontekstem tuż przed ostateczną generacją odpowiedzi. Jego siła polega na tym, że model nie musi już pamiętać wszystkiego, bo dostaje to, co trzeba, dokładnie wtedy, gdy trzeba; niemniej słabość tego układu objawia się w momencie, gdy podamy maszynie zły kontekst. AI nie znosi architektury, ona ją dopiero brutalnie wymusza.

W codziennej praktyce inżynieryjnej RAG składa się z kilku etapów, które mają charakter tyleż techniczny, co wręcz instytucjonalny, wymagając od nas niemal biurokratycznej dyscypliny w zarządzaniu danymi. Najpierw dokument musi zostać poddany rygorystycznej obróbce, co oznacza czyszczenie, wydzielanie sensownych fragmentów oraz precyzyjne opisywanie ich metadanymi w indeksie. Następnie zapytanie użytkownika podlega reprezentacji wektorowej, po czym system wyszukuje najbardziej trafne fragmenty, ewentualnie je rerankuje i łączy z instrukcją dla modelu. Jeżeli w tym precyzyjnym łańcuchu zawiedzie jakość podziału tekstu, selekcja kontekstu lub polityka uprawnień, cała obietnica inteligentnej wiedzy redukuje się do eleganckiej odmiany chaosu. Wiele ambitnych projektów przegrywa właśnie na tym etapie, bowiem ich twórcy zapominają, że ostateczny sukces zależy od higieny wejścia, a nie od samego modelu. W nauce byłoby to śmieszne, w biznesie bywa po prostu bardzo drogie, a w prawie kończy się sporem.

Dlatego właśnie musimy osobno zatrzymać się przy pojęciu chunku, czyli fragmentu tekstu stanowiącego podstawową jednostkę informacyjną, który bywa używany tak lekkomyślnie, jakby chodziło o zwykły ścinek. Tymczasem chunk jest jednostką trudnego kompromisu między

semantyką, ograniczeniem okna kontekstowego, kosztami wyszukiwania oraz ryzykiem bolesnej utraty sensu w procesie przetwarzania. Jeśli fragment jest zbyt mały, proces wyszukiwania zwraca jedynie okruchy znaczenia pozbawione jakiejkolwiek ramy argumentacyjnej, co czyni odpowiedź modelu powierzchowną i często błędną. Jeśli natomiast jest zbyt duży, rośnie szum informacyjny oraz prawdopodobieństwo, że do promptu trafi materiał skutecznie rozmywający właściwą odpowiedź na zadane pytanie. Dokumentacja LangChain konsekwentnie rekomenduje RecursiveCharacterTextSplitter jako domyślną strategię, ponieważ próbuje ona zachowywać akapity i zdania, dbając o semantyczną spójność elementów. To oznacza, że chunk nie jest technicznym odpadem, lecz formą redakcji poznawczej, będąc w istocie małą konstytucją kontekstu.

Warto dodać, że chunk pełni w nowoczesnych systemach podwójną rolę, będąc jednocześnie porcją wiedzy dla algorytmu i porcją doświadczenia dla człowieka w procesie interakcji. W zaawansowanych aplikacjach tekst oraz elementy interfejsu są dostarczane stopniowo, przy czym strumieniowanie, czyli praktyka przesyłania danych w czasie rzeczywistym, pozwala skrócić psychologiczny czas oczekiwania użytkownika na finalny rezultat. Vercel AI SDK traktuje ten proces jako element pierwszego rzędu, a nie jedynie estetyczny ozdobnik mający maskować opóźnienia wynikające z natury modeli. Stąd bierze się swoisty paradoks naszych czasów, w których systemy ocenia się nie po tym, czy wiedzą najwięcej, lecz po płynności ich cyfrowego współdziałania. Najbardziej wyrafinowane architektury potrafią dać użytkownikowi kojące odczucie współpracy z maszyną, która wprawdzie myśli na raty, ale nie każe nam oglądać własnej zadyszki. Postęp bywa komiczny, zwłaszcza gdy jego miarą staje się sztuka maskowania technicznych niedoskonałości.

W tym miejscu do gry wchodzi agent, czyli autonomiczny system zdolny do podejmowania decyzji i wykonywania zadań – pojęcie, które zostało na rynku jednocześnie przedwcześnie zużyte i niedomyślane. Agentem warto nazywać dopiero taki system, który potrafi samodzielnie przyjąć cel, rozłożyć go na kroki, dobrać narzędzia i kontrolować wynik w pętli decyzyjnej. W tym znaczeniu nie jest on jedynie rozszerzonym chatbotem, lecz minimalną jednostką delegowanej sprawczości, zmieniającą ontologię oprogramowania z prostych narzędzi w aktywne środowiska operacyjne. Despoudis wydobywa tu kategorię agenta racjonalnego, wiążąc ją z nową logiką władzy cyfrowej i programowalną alokacją uwagi oraz działania. Kto projektuje funkcję celu, ten po cichu rozstrzyga, które stany świata będą przez maszynę promowane jako pożądane, co ma fundamentalne skutki ekonomiczne. Z punktu widzenia prawa to pytanie o rozkład odpowiedzialności między projektanta, operatora oraz użytkownika delegującego intencję maszynie.

Tu znów przydaje się anegdota, bowiem dawna aplikacja była jak urzędnik przy okienku, przy którym przynajmniej wiedzieliśmy, że trzeba nacisnąć guzik i czekać na pieczątkę. Agent przypomina raczej nadgorliwego asystenta, który mówi, że już się wszystkim zajął, po czym w połowie przypadków rzeczywiście rozwiązuje nasz problem z godną podziwu sprawnością. W drugiej połowie potrafi jednak zamówić dwa noclegi w różnych miastach, ponieważ wykrył w danych rzekomą optymalizację elastyczności pobytu, której nikt od niego nie oczekiwał. Absurd tej sytuacji nie bierze się z tego, że maszyna jest głupia, lecz z faktu, że jest ona przerażająco i bezrefleksyjnie wykonawcza. Cywilizacja przez stulecia uczyła się żyć z błędnymi opiniami, niemniej znacznie gorzej radzi sobie z błędami operacyjnymi wykonywanymi z uprzejmą, cyfrową automatyzacją. Cywilizacja zbudowała lokaja, a potem musiała mu tłumaczyć, że nie każda karteczka wsunięta pod drzwi jest instrukcją od właściciela domu.

MCP: Nowy standard integracji czy nowe wyzwanie dla bezpieczeństwa?

Droga do Model Context Protocol, czyli otwartego standardu bezpiecznych połączeń między modelami a danymi, była wybrukowana kosztownym chaosem integracyjnym. Rynek wreszcie zrozumiał, że zmuszanie każdej aplikacji AI do osobnej, ręcznej integracji z każdym narzędziem to przepis na inżynieryjny koszmar. Powstawał klasyczny problem mnożenia połączeń, gdzie wysokie koszty utrzymania szły w parze z nierówną jakością zabezpieczeń i permanentnym uwięzieniem u jednego dostawcy, znanym jako vendor lock-in. Anthropic, inicjując ten standard, nakreślił wizję bezpiecznego pomostu, a OpenAI szybko podchwyciło temat w dokumentacji Responses API, pokazując obsługę konektorów i zdalnych serwerów. W tej nowej architekturze wywoływanie narzędzi może dziać się automatycznie lub zależeć od jawnej, świadomej zgody dewelopera. To moment istotny cywilizacyjnie, ponieważ integracja przestaje być rzemieślniczym, ręcznie szytym wyjątkiem, a staje się po prostu kolejną warstwą protokołu.

MCP ma znaczenie znacznie głębsze niż zwykła wygoda programistyczna, ponieważ ustanawia nowy, racjonalny kompromis między uniwersalnością modelu a lokalnością narzędzi. Model nie musi już posiadać wbudowanej, specjalistycznej logiki dla każdego systemu, z którym przyjdzie mu współpracować. Wystarczy, że potrafi on sprawnie odkryć zasoby serwera, dostępne prompty oraz rygorystyczne warunki autoryzacji, na jakich wolno mu wywołać konkretne funkcje. Dzięki temu możliwa staje się modularna architektura środowiska pracy AI, co drastycznie obniża koszty integracji w skali makro. Jest to kluczowe z punktu widzenia prawa, gdyż pomaga oddzielić jądro modelu od wrażliwych sekretów biznesowych i logiki operacyjnej. Wreszcie, zmienia to kulturę

techniczną, przesuwając prestiż z popisowej „inteligencji modelu” na rzetelny projekt ekosystemu współdziałania. AI nie znosi architektury. Ona ją dopiero wymusza.

Warto jednak zachować zdrowy sceptycyzm, gdyż popularność standardu nigdy nie gwarantuje dojrzałości całego otoczenia technologicznego. Oficjalna dokumentacja OpenAI wyraźnie ostrzega przed bezkrytycznym używaniem adresów URL zwracanych przez zewnętrzne konektory i serwery MCP, które mogą być niepewne. Zaleca się raczej korzystanie z oficjalnych serwerów utrzymywanych przez sprawdzonych dostawców usług, co stanowi sygnał o fundamentalnym znaczeniu dla bezpieczeństwa danych. Wraz ze standaryzacją nie znikają przecież odwieczne problemy zaufania, integralności odpowiedzi czy podatności na ataki; one jedynie zmieniają swój wektor. Protokół nie jest ostatecznym zbawieniem ani technologicznym panaceum, lecz ramą, w której można projektować systemy albo genialnie, albo fatalnie. Teraz jednak możemy to robić ze znacznie większym rozmachem i systemową pewnością.

Pierwsza faza zachwytu nad sztuczną inteligencją skupiała się na pytaniu, czy model umie mówić, natomiast faza obecna dotyczy kwestii znacznie mniej romantycznej. Brzmi ona: czy wolno mu działać w świecie rzeczywistym? To pytanie nie jest jedynie ozdobą debaty, lecz stanowi jej oś, wokół której kręci się przyszłość systemów agentowych, czyli autonomicznych bytów zdolnych do podejmowania decyzji. W tym nowym świecie bezpieczeństwo nie jest opcjonalnym dodatkiem do funkcjonalności, lecz warunkiem jej cywilizacyjnej dopuszczalności w obrocie gospodarczym. Im bardziej model przechodzi od generowania zdań do wykonywania czynności, tym bardziej architektura systemu zaczyna przypominać konstrukcję odpowiedzialności publicznej. Musimy wiedzieć, kto uruchamia narzędzie, na jakiej podstawie i w jakim reżimie audytu, co wykracza daleko poza ramy czystej informatyki. To ekonomia ryzyka, prawo kompetencji i antropologia zaufania w jednym.

Właśnie w tym miejscu rozstrzyga się przewaga dojrzałych systemów nad jarmarcznym technoentuzjazmem, który karmi się obietnicami bez pokrycia. OpenAI w swojej dokumentacji narzędzi pokazuje, że modele mogą korzystać z wyszukiwania, plików i zdalnych serwerów, ale robi to z ogromną dozą inżynieryjnej ostrożności. Oficjalne materiały ostrzegają, by nie ufać bezkrytycznie odpowiedziom pochodzącym z niezweryfikowanych źródeł i zawsze stosować rygorystyczną kontrolę po stronie aplikacji. To bardzo istotny sygnał dla projektantów, że samo rozszerzenie sprawczości modelu nie jest jeszcze sukcesem, a jedynie otwarciem nowej puszki Pandory. Prawdziwym osiągnięciem jest dopiero takie zwiększenie możliwości, które pozostaje pod ścisłą kontrolą zasad, logów i autoryzacji. Sukcesem jest sprawczość ujęta w karby procedur. Reszta jest marketingiem.

W świetle tych faktów najlepiej widać, jak bardzo myląca bywa popularna opowieść o AI jako o genialnym asystencie. Genialny asystent jest figurą literacką, podczas gdy w środowisku produkcyjnym mamy do czynienia z bytem, który jest inteligentnym, lecz podatnym na manipulację wykonawcą. Prompt injection, czyli technika manipulowania modelem poprzez złośliwe instrukcje ukryte w danych, pozostaje problemem, który ewoluuje wraz z możliwościami agentów. Nowoczesne podejście do projektowania nie zakłada absolutnego wykrycia każdego ataku, lecz skupia się na ograniczaniu skutków ewentualnej manipulacji, gdy ta już nastąpi. OWASP ujmuje rzecz brutalnie: problemem nie jest tylko to, co użytkownik powiedział agentowi wprost. Problemem jest także to, co z zewnątrz próbuje powiedzieć agentowi świat, dążąc do eskalacji uprawnień i wycieku danych.

W tym miejscu humor absurdu przychodzi niemal sam, obnażając paradoksy naszego pędu ku nowoczesności i cyfrowej wygodzie. Przez lata uczono obywatela internetu, żeby nie klikał w podejrzane linki i nie wierzył egzotycznym darczyńcom pilnie potrzebującym danych karty kredytowej. Tymczasem teraz z dumą produkujemy systemy, którym powierzamy czytanie całego internetu w naszym imieniu. Człowiek został wreszcie odciążony z nudnego obowiązku czujności, po czym odkrył, że musi nauczyć tej czujności maszynę czytającą wszystko z miną pilnego archiwisty. Dawniej naiwny był użytkownik, dziś naiwny bywa agent, co pokazuje, że postęp bywa komiczny w swojej przewrotności. Cywilizacja zbudowała lokaja, a potem musiała mu tłumaczyć, że nie każda karteczka wsunięta pod drzwi jest instrukcją od właściciela domu.

Architektura agentów: Bezpieczeństwo, koszty i kontrola jakości

Współczesna inżynieria uczy nas, że traktowanie backendu jako bezpiecznego pośrednika jest aksjomatem, toteż klucze API nie mogą znaleźć się w kodzie klienta. Każdy publiczny punkt styku z systemem wymaga rygorystycznej walidacji wejścia oraz precyzyjnej kontroli natężenia ruchu, co chroni infrastrukturę przed niekontrolowanym przeciążeniem. Dokumentacja OpenAI wskazuje, że restrykcje dotyczą liczby żądań oraz objętości tokenów, przy czym błąd 429 sygnalizuje zazwyczaj przekroczenie limitów planu. W praktyce oznacza to, że architektura aplikacji AI musi być ekonomicznie świadoma, bowiem każde zapytanie to nie tylko akt komunikacji, lecz konkretny koszt i latencja. Rate limiting, czyli technika ograniczania liczby operacji w jednostce czasu, pozwala na utrzymanie sensownego rachunku operacyjnego. AI nie znosi architektury, ona ją dopiero wymusza.

Klasyczny ekonomista dostrzegłby w tym miejscu oczywistość, bowiem modele językowe drastycznie obniżają koszty syntezy informacji, lecz równocześnie generują zupełnie nowe kategorie wydatków transakcyjnych. Ciężar finansowy przesuwa się nieuchronnie w stronę kosztów kontroli oraz stałego monitoringu systemów. Wszak wdrażanie agenta, czyli autonomicznego systemu zdolnego do podejmowania decyzji w imieniu użytkownika, przypomina fabrykę bez kontroli jakości. Taki model biznesowy nie jest wcale innowacyjny, lecz staje się po prostu wydajniejszą maszyną do generowania szybkich strat. W nauce byłoby to śmieszne, w biznesie bywa natomiast zabójczo drogie. Prawdziwy koszt AI leży w inwestycjach niezbędnych do tego, by sprawczość modelu stała się prawnie, organizacyjnie i ekonomicznie znośna. To prowadzi nas nieuchronnie w objęcia paragrafów, gdzie kończy się radosna twórczość, a zaczyna twarda odpowiedzialność.

Prawne rygory i techniczna odpowiedzialność w erze agentów

W Europie nie da się już uczciwie mówić o projektowaniu systemów AI bez uwzględnienia AI Act, czyli europejskiego rozporządzenia regulującego projektowanie i wdrażanie systemów sztucznej inteligencji, które przestało być jedynie straszakiem dla teoretyków. Komisja Europejska wskazuje wyraźnie, że Akt o AI wszedł w życie 1 sierpnia 2024 roku i choć w pełni zacznie być stosowany od 2 sierpnia 2026 roku, to zegar tyka znacznie szybciej, niż wydaje się to optymistom. Zakazane praktyki oraz obowiązki w zakresie AI literacy, czyli budowania kompetencji w rozumieniu i krytycznym wykorzystywaniu technologii, zaczęły obowiązywać już 2 lutego 2025 roku, a rygory dla modeli ogólnego przeznaczenia weszły w życie w sierpniu tego samego roku. Oznacza to, że problem zgodności prawnej nie jest mglistą wizją odległej przyszłości, lecz realnym, trwającym procesem, który redefiniuje zasady gry. Systemy używane do rekrutacji, oceny dostępu do usług czy zarządzania pracownikami nie mogą być już traktowane jak niewinne eksperymenty produktowe, bo wchodzą w pole prawne o rosnącej gęstości i surowości.

Z perspektywy budowania aplikacji AI ma to konsekwencje niezwykle konkretne, wręcz bolesne dla tych, którzy przywykli do technologicznego „Dzikiego Zachodu”. Jeśli konstruujemy agenta rekrutacyjnego, finansowego czy medycznego, nie wystarczy już, że model mówi przekonująco, szybko i z wdziękiem cyfrowego erudyty. Twórca musi być w stanie wykazać, jakie dane system przetwarza, jakie ryzyka generuje dla użytkownika końcowego oraz jak w praktyce wygląda nadzór człowieka nad autonomicznym procesem. Europejski reżim prawny nie pyta z zachwytem, czy system jest nowatorski lub czy jego architektura budzi podziw na konferencjach w Dolinie Krzemowej. Pyta chłodno i bez emocji, czy jest on dopuszczalny, rozliczalny i czy w swoim dążeniu

do optymalizacji nie narusza praw podstawowych jednostki. I bardzo dobrze, bo kultura techniczna, która nie znosi zewnętrznego testu legalności, najczęściej nie jest kulturą odwagi, lecz kulturą wygodnego infantylizmu.

W cieniu tych regulacji pozostaje kwestia prywatności, która w systemach opartych na danych staje się problemem egzystencjalnym. W architekturach typu RAG, czyli rozwiązaniach łączących generowanie tekstu z dynamicznym wyszukiwaniem informacji z zewnętrznych źródeł, problem ochrony danych osobowych nie jest marginalnym przypisem do dokumentacji. Jest on centralnym punktem zapalnym, ponieważ wysyłanie treści zawierających identyfikatory osobowe czy wrażliwe informacje kontraktowe do zewnętrznych modeli wymaga czegoś więcej niż tylko polityk dostępu. Niezbędna staje się rygorystyczna anonimizacja, minimalizacja danych oraz niezwykle precyzyjna selekcja kontekstu, który trafia do modelu. Im sprawniejszy jest proces wyszukiwania, tym większa staje się pokusa, by do indeksu wrzucać wszystko, co wpadnie nam w ręce, w nadziei na lepszą jakość odpowiedzi. To pokusa niebezpieczna, gdyż RAG bez rygoru danych może stać się maszyną do błyskawicznego odnajdywania rzeczy, których model w ogóle nie powinien widzieć. Techniczna elegancja nie unieważnia obowiązku prawnego; przeciwnie, często go zaostrza, zwiększając skalę możliwego wycieku i tempo reprodukcji błędu.

Na tym tle widać również, jak głęboko mylone bywa pojęcie jakości w systemach opartych na osadzeniach, czyli geometryzacji podobieństwa semantycznego poprzez mapowanie danych do przestrzeni wektorowej. Jakość nie polega wyłącznie na tym, że odpowiedź brzmi mądrze i jest sformułowana poprawną polszczyzną, lecz na tym, że trafność wyszukania i granularność fragmentów tworzą architekturę relewancji, która nie rozsadza bezpieczeństwa. OpenAI wskazuje, że wyszukiwanie opiera się na bazach wektorowych, a mechanizmy przeszukiwania plików pozwalają ograniczać liczbę zwracanych wyników, redukując użycie tokenów kosztem możliwego spadku jakości odpowiedzi. Ten kompromis jest znakomitym przykładem ogólniejszej reguły: w systemach AI nie istnieje darmowa pełnia, a każda poprawa w jednym wymiarze wymusza zapłatę w innym. Więcej kontekstu może oznaczać większy koszt i ryzyko szumu, podczas gdy mniej kontekstu to większa szybkość, ale słabsze uziemienie odpowiedzi w faktach. Inżynieria AI jest więc sztuką świadomego handlu ograniczeniami, a nie prostą drogą do technologicznej wszechmocy.

Zresztą sama idea osadzeń wektorowych również dojrzewa, wychodząc poza proste ramy tekstowe. Nadal obowiązuje podstawowy paradygmat, w którym osadzenia służą wyszukiwaniu, rekomendacjom i klasyfikacji, a podobieństwo cosinusowe pozostaje standardową miarą podobieństwa w tej matematycznej przestrzeni. Jednocześnie rynek gwałtownie przesuwa się w stronę multimodalności, co zmienia zasady gry dla wszystkich projektantów systemów agentowych. Google ogłosiło w marcu 2026 roku model Gemini Embedding 2 jako natywnie multimodalny,

zdolny do mapowania tekstu, obrazów, wideo i audio do jednej, wspólnej przestrzeni osadzeń. To ważny sygnał, bo oznacza, że przyszłe systemy RAG nie będą już działać wyłącznie na korpusach tekstowych, lecz zaczną poruszać się po środowiskach mieszanych. W takim świecie dokument, wykres, nagranie ze spotkania i obraz mogą współtworzyć jeden krajobraz semantyczny, dając agentom niespotykany dotąd wgląd w kontekst. Zysk poznawczy jest tu ogromny, ale ryzyko rośnie proporcjonalnie, bo drastycznie wzrasta złożoność walidacji takich systemów i trudność ich audytu. AI nie znosi architektury; ona ją dopiero wymusza, a my musimy nauczyć się w niej żyć.

Od magii modeli do konstytucji operacyjnej: nowa logika sprawczości

Intuicje zawarte w materiałach Fundacji Dobre Państwo zasługują na coś więcej niż tylko uprzejme skinienie głową ze strony branży technologicznej. Gdy tamtejsze analizy opisują agenta racjonalnego jako nową logikę władzy w gospodarce cyfrowej, uderzają one w sam środek tarczy. Dzisiejszy spór o sztuczną inteligencję rzadko dotyczy tego, czy modele będą mądrzejsze, co wydaje się już niemal przesądzone. Spór ten dotyczy raczej tego, kto i na jakich zasadach będzie organizował ich cele oraz hierarchię błędów tolerowalnych. To już nie jest zagadnienie marginalne ani wyłącznie techniczne, lecz kwestia ustrojowa w najszerszym tego słowa znaczeniu.

Najważniejsza teza dzisiejszej debaty brzmi zatem następująco: przyszłość aplikacji AI nie należy do tych, którzy najgłośniej opowiadają o magii wielkich modeli. Należy ona do projektantów potrafiących zbudować rygorystyczny porządek operacyjny dla delegowanej sprawczości. Prompt staje się konstytucją zadania, a embeddings, czyli geometryzacja podobieństwa semantycznego poprzez mapowanie danych do przestrzeni wektorowej, wyznaczają ramy sensu. RAG, czyli architektura łącząca generowanie tekstu z dynamicznym wyszukiwaniem informacji z zewnętrznych źródeł, służy dozbrajaniu modelu w relewantny kontekst. MCP, czyli otwarty standard bezpiecznych połączeń między modelami a narzędziami, staje się protokołem interoperacyjnej sprawczości. W takim układzie bezpieczeństwo przestaje być działem wsparcia, a staje się ontologiczną zasadą całego systemu.

Istnieje pewna anegdota, być może zbyt prawdziwa, by uznać ją za żart, o korporacyjnym marzeniu o agencie, który „sam coś załatwi”. Jest to wizja pociągająca, niemniej niemal nikt nie marzy równie intensywnie o logach, politykach dostępu czy testach odporności na prompt injection. Wszyscy chcą dziś cyfrowego ministra sprawczości, ale mało kto ma ochotę napisać mu rzetelną konstytucję. Bez niej jednak każdy genialny minister szybko zamienia się w kłopotliwego kuzyna historii, generując

więcej chaosu niż pożytku. W nauce byłoby to śmieszne, w biznesie bywa drogie, a w prawie zazwyczaj kończy się bolesnym sporem.

Najdonioślejsza zmiana w architekturze cyfrowej nie polega na tym, że modele językowe sprawniej budują zdania podrzędnie złożone. Polega ona na przesunięciu całego środowiska projektowania od logiki interfejsu ku logice środowiska działania. Dawna aplikacja była przede wszystkim miejscem, w którym użytkownik wykonywał serię jawnych i przewidywalnych czynności. Dzisiejsza aplikacja AI staje się systemem, w którym człowiek ustanawia intencję, a resztę przejmuje warstwa pośrednia złożona z modeli i protokołów. Właśnie dlatego Responses API jest dziś promowane przez OpenAI jako fundament nowych projektów, oferujący znacznie bardziej „agentowe" możliwości niż starsze podejścia. Planowane na sierpień 2026 roku usunięcie Assistants API to jasny sygnał, że rynek porządkuje się wokół zadań, a nie tylko rozmów.

W tym sensie agent nie jest jedynie efektownym dodatkiem, lecz figurą centralną nowego układu gospodarczego. Z perspektywy ekonomii zmienia on sposób organizacji kosztów transakcyjnych poprzez agregację rozproszonych mikroczynności. Tradycyjnie użytkownik sam musiał przedzierać się przez formularze, podczas gdy agent przejmuje tę pracę koordynacyjną na siebie. Jednak ta wygoda wprowadza nowe koszty, które musimy zapłacić w walucie nadzoru, walidacji oraz zgodności prawnej. Korzyść nie polega więc na magicznej eliminacji wydatków, lecz na ich głębokiej redystrybucji. Despoudis zauważa, że agent racjonalny to nowa logika organizacji decyzji, a nie tylko cyfrowy ozdobnik.

Ta nowa logika posiada zarazem wymiar antropologiczny, który wykracza poza zwykłą użyteczność oprogramowania. Przez dekady sieć uczyła nas, że sprawczość polega na świadomym klikaniu w przyciski, będące świecką wersją dźwigni woli. Aplikacja agentowa rozsadza ten porządek, przestając pytać użytkownika o każdy mikrogest, a zamiast tego prosząc o cel i granice działania. Wchodzimy tym samym w relację quasi-pełnomocnictwa, co z punktu widzenia psychologii nie jest sytuacją banalną. Człowiek znacznie łatwiej wybacza maszynie błąd informacyjny niż błąd wykonawczy, taki jak błędny przelew czy niefortunna decyzja personalna. Gdy AI przechodzi od opisu świata do ingerencji w świat, moralna temperatura całego systemu gwałtownie rośnie.

Architektura suwerenności: Standardy, agenty i nowa ekonomia bram

Na tym tle szczególnego znaczenia nabiera kwestia warstw abstrakcji, które w świecie oprogramowania pełnią rolę podobną do protokołów dyplomatycznych w podzielonym świecie.

Historycznie rzecz biorąc, każdy dostawca modeli językowych oferował własne API, unikalną strukturę odpowiedzi oraz specyficzny sposób pracy z narzędziami, co tworzyło cyfrowe lenna. To rodziło klasyczny problem vendor lock-in, czyli sytuację, w której techniczna wierność jednemu dostawcy staje się pułapką, z której ucieczka kosztuje więcej niż sam produkt. Kto raz głęboko zintegrował swój system z konkretnym modelem, ten płacił później wysoką cenę za każdą próbę migracji do konkurencji. Współczesne zestawy narzędzi programistycznych, znane jako SDK, próbują ten problem osłabić poprzez wprowadzanie standardów pośredniczących. Vercel AI SDK jawnie pozycjonuje się jako warstwa ujednolicająca dostęp do wielu modeli, oferując funkcje generowania, strumieniowania oraz produkcji danych ustrukturyzowanych w sposób niezależny od dostawcy. Jego znaczenie nie polega wyłącznie na wygodzie pisania kodu, lecz na obniżaniu kosztu strategicznej zależności, co w biznesie bywa drogie, a w architekturze systemów krytyczne.

Jeżeli można przełączyć model bez rozrywania całej logiki aplikacji, zyskuje się nie tylko elastyczność techniczną, lecz także realną przewagę negocjacyjną w rozmowach z gigantami technologicznymi. W ekonomii platform jest to różnica fundamentalna, definiująca podmiotowość gracza na rynku. Kto ma realną możliwość wyjścia z ekosystemu, ten ma większą sprawczość niż ten, kto jedynie liczy na łaskę aktualnego dostawcy i jego cennika. Oczywiście warstwa abstrakcji nie jest panaceum na wszystkie bolączki integracyjne, ponieważ ujednolicony interfejs bywa okupiony utratą dostępu do pewnych unikalnych możliwości. Powstaje więc napięcie stare jak sama gospodarka techniczna, gdzie uniwersalność walczy o lepsze z wyspecjalizowaną wydajnością. Z jednej strony pragniemy wymienialności komponentów, z drugiej zaś najbardziej zaawansowane funkcje zwykle pojawiają się jako lokalne, specyficzne dla dostawcy przewagi, których nie da się łatwo zamknąć w standardzie.

Dlatego dojrzała architektura nie powinna wpadać ani w dogmat pełnej abstrakcji, ani w dogmat natywnej wierności jednemu dostawcy, lecz operować na wielu poziomach. Trzeba projektować warstwowo, oddzielając to, co zmienne, od tego, co stanowi trwały fundament systemu. To, co stanowi rdzeń biznesowej logiki, warto utrzymywać w formie możliwie przenaszalnej, chroniąc intelektualny kapitał firmy przed kaprysami rynku. Z kolei to, co daje unikatową przewagę i ma uzasadniony zwrot z inwestycji, można wiązać mocniej z konkretnym dostawcą, akceptując ryzyko w zamian za wydajność. To nie jest jedynie kompromis estetyczny czy inżynierska elegancja, lecz chłodny rachunek suwerenności technologicznej w niepewnych czasach. Podobnie należy spojrzeć na relację między natywnymi API, SDK i frameworkami orkiestracyjnymi, które tworzą hierarchię kontroli nad procesem przetwarzania danych.

Natywne API daje największą kontrolę i najszybszy dostęp do najnowszych funkcji, ale wymaga od programisty rzemieślniczej precyzji w każdym detalu. SDK upraszcza codzienną pracę i znacząco

zmniejsza koszt integracji, podczas gdy frameworki takie jak LangChain służą porządkowaniu bardziej złożonych przepływów pracy. W takich potokach operacyjnych retrieval, czyli proces odzyskiwania informacji, pamięć oraz wieloetapowe rozumowanie muszą współpracować w idealnej harmonii. LangChain nadal eksponuje strategie dzielenia tekstu, w tym RecursiveCharacterTextSplitter, czyli narzędzie do cięcia danych na fragmenty, które model jest w stanie przetworzyć bez utraty sensu. Praktyka pokazuje bowiem, że nawet najbardziej błyskotliwy model polegnie, jeśli dostanie źle przygotowany materiał wejściowy, co obnaża naiwność wiary w czystą inteligencję bez struktury. Przemysł AI powoli dojrzewa do przykrej, lecz zbawiennej prawdy: spektakularna moc modelu nie znosi potrzeby rzemiosła integracyjnego, lecz wręcz ją potęguje.

Im potężniejszy model, tym bardziej kompromitujące stają się banalne błędy w pipeline, co prowadzi nas prosto do Model Context Protocol (MCP). Warto zatrzymać się przy tym standardzie, ponieważ właśnie on może stać się najbardziej brzemiennym w skutki rozwiązaniem technologicznym tej dekady. Anthropic opisał MCP jako otwarty standard łączenia asystentów AI z narzędziami i danymi, co ma szansę zakończyć erę chaotycznych, ręcznych integracji. OpenAI potwierdziło praktycznie, że ich Responses API współpracuje ze zdalnymi serwerami MCP, co pozwala modelowi na dynamiczne importowanie listy narzędzi w toku generowania odpowiedzi. Jeżeli ten ekosystem dojrzeje, aplikacja przestanie być zbiorem dziesiątek zamkniętych API, które programista musi mozolnie łączyć w jedną, kruchą całość. Zacznie raczej operować w pejzażu usług wystawiających swoje możliwości w protokole zrozumiałym dla wielu agentów i wielu modeli jednocześnie.

To jest początek zupełnie nowego rynku interoperacyjności, czyli zdolności systemów do wymiany danych, gdzie nie handluje się już dostępem do stron, lecz konkretnymi zdolnościami operacyjnymi maszyn. Konsekwencje tego przesunięcia są dalekosiężne i dotykają samej definicji tego, czym jest współczesne oprogramowanie użytkowe. Dziś jeszcze myślimy o aplikacji jako o obiekcie zamkniętym, który posiada własny frontend, menu i ściśle zaprojektowaną ścieżkę użytkownika. W świecie agentowym aplikacja coraz częściej staje się raczej wystawcą możliwości niż teatralną sceną dla ludzkich kliknięć i przesunięć kursora. To, co dawniej było zaprojektowane wyłącznie dla ludzkiego oka, musi być teraz zaprojektowane również dla rozumu maszynowego, co wymusza nową ontologię interfejsów. Narzędzie powinno dać się odkryć przez algorytm, a każda operacja musi posiadać precyzyjny opis, schemat oraz jasną politykę autoryzacji.

Zasób cyfrowy powinien być jednoznacznie identyfikowalny, a prompt, czyli instrukcja sterująca modelem, musi dać się osadzić w rygorze wykonawczym bez marginesu na błąd. Można zaryzykować tezę mocniejszą: przyszłość sieci nie należy do tych, którzy zbudują najpiękniejsze

ekrany, lecz do tych, którzy najlepiej opiszą swoje zdolności dla agentów. Brzmi to może sucho i technokratycznie, ale w gruncie rzeczy jest to proces głęboko polityczny, dotykający fundamentów władzy w cyfrowym świecie. Kto kontroluje protokoły dostępu do działania, ten kontroluje nową infrastrukturę pośrednictwa, stając się strażnikiem bram w nowej gospodarce. W klasycznym internecie dominację zdobywało się przez uwagę użytkownika i zamknięcie go w złotym więzieniu własnego ekosystemu. W internecie agentowym przewaga przesunie się ku temu, kto kontroluje przejścia między intencją użytkownika a skutecznym wykonaniem zadania przez maszynę.

To oznacza, że walka o standardy stanie się walką o nowe formy renty ekonomicznej, gdzie zysk płynie z samej obecności na styku intencji i realizacji. Już dziś widać zalążki tej logiki w praktykach integracyjnych, w katalogach narzędzi oraz w całej rosnącej gospodarce typu "agent ready". To nie jest tylko kwestia inżynierii oprogramowania, lecz nowa ekonomia bram, w której stawką jest kontrola nad przepływem sprawstwa w sieci. Na tym tle materiały fundacji Dobre Państwo zyskują dodatkową ostrość, przypominając, że technologia nigdy nie rozwija się w próżni społecznej. Tekst o epistemologii postępu argumentuje, że postęp technologiczny bez rozwoju instytucjonalnego grozi jedynie pogłębieniem asymetrii władzy między twórcami a użytkownikami. W połączeniu z analizą agenta racjonalnego daje to silny wniosek: agenty nie są politycznie niewinne, lecz niosą w sobie wartości swoich stwórców.

To, co agenty mają optymalizować, kto ustala ich funkcję celu oraz jakie błędy uznaje się za akceptowalne, nie jest technicznym detalem, lecz sporem o architekturę porządku społecznego. I tu trzeba zaatakować konformizm, który każe nam wierzyć, że technologia jest siłą natury, na którą nie mamy żadnego wpływu. Najbardziej infantylna wersja debaty o sztucznej inteligencji głosi, że technologia sama "zdecyduje", dokąd zmierza świat, co jest wygodną bajką dla unikających odpowiedzialności. Technologia nie posiada własnej teleologii, czyli wewnętrznego celu, do którego dąży niezależnie od ludzkiej woli i projektów. Ma za to sponsorów, architektów, regulatorów oraz konkretne budżety, które determinują jej kształt i sposób oddziaływania na rzeczywistość. Agent nie jest duchem dziejów ani metafizyczną siłą, lecz implementacją określonych preferencji w konkretnym otoczeniu gospodarczym i prawnym.

Kto mówi inaczej, ten najczęściej sprzedaje mgłę albo kupuje sobie moralną ulgę, uciekając od trudnych pytań o sprawiedliwość i kontrolę. AI nie znosi architektury. Ona ją dopiero wymusza, stawiając nas przed koniecznością zdefiniowania nowych zasad współistnienia ludzi i maszyn. W nauce byłoby to śmieszne, gdybyśmy udawali, że narzędzia same się tworzą, ale w biznesie takie podejście bywa niezwykle kosztowne. Cywilizacja zbudowała lokaja, a potem musiała mu tłumaczyć, że nie każda karteczka wsunięta pod drzwi jest instrukcją od prawowitego właściciela domu. Musimy zatem projektować systemy z pełną świadomością ich politycznego ciężaru, bo

każda linia kodu w standardzie takim jak MCP jest małą konstytucją przyszłego porządku. Ostatecznie suwerenność technologiczna to nie tylko posiadanie własnych serwerów, ale przede wszystkim zdolność do decydowania o regułach, według których te serwery ze sobą rozmawiają.

Od feudalizmu formularzy do konstytucji współpracy

Pozwólmy sobie na chwilę anegdoty, gdyż nawet w świecie surowych protokołów warto wszak zachować przestrzeń na odrobinę ironicznego uśmiechu. Wyobraźmy sobie klasyczną witrynę internetową, która z niemal barokową dumą prezentuje użytkownikowi labirynt siedmiu zakładek, dziewięciu migających banerów oraz formularz tak absurdalnie długi, że jego wypełnienie sugeruje ubieganie się o obywatelstwo nowego państwa. Ta cyfrowa architektura była wprawdzie męcząca, lecz przynajmniej uczciwie i bez ogródek komunikowała, że zamierza bezpowrotnie skraść nasz cenny czas. Współczesna aplikacja agentowa, czyli autonomiczny system zdolny do podejmowania decyzji w imieniu użytkownika, działa znacznie subtelniej, kusząc nas obietnicą całkowitego zwolnienia z wysiłku. Cywilizacja płynnie przeszła zatem od feudalizmu uciążliwych formularzy do specyficznego, oświeconego absolutyzmu delegacji. To niewątpliwy postęp, niemniej warto pamiętać, że każdy absolutyzm pozostaje formą tyranii, dopóki nie zostanie mu narzucona surowa konstytucja.

Dlatego właśnie nadchodząca przyszłość sieci będzie coraz mniej przypominała chaotyczną kolekcję stron, a bardziej precyzyjnie negocjowaną umowę społeczną między człowiekiem, modelem i instytucją. W tym nowym układzie użytkownik wnosi suwerenną intencję oraz świadomą zgodę, podczas gdy model dostarcza unikalną zdolność wnioskowania i sprawnej kompozycji działań. Instytucja bowiem musi zapewnić sztywne ramy dostępu, politykę odpowiedzialności oraz niezbędne warunki legalności całego procesu. Każda epoka ma swoją naiwną metaforę technologiczną, lecz tutaj stawką jest realna architektura współpracy, a nie tylko estetyka interfejsu. Tam, gdzie ten trójkąt kompetencji zostanie zaprojektowany z należytą mądrością, otrzymamy systemy autentycznie użyteczne i bezpieczne. Jeśli jednak którakolwiek ze stron uzna się za zbędną, otrzymamy technokratyczny bałagan lub kosztowną zabawkę; reszta jest marketingiem.

Koniec ery promptowania: Architektura kontekstu i kontrola

Każda epoka ma swoją naiwną metaforę technologiczną, a w dobie wielkich modeli językowych jest nią utożsamienie dostępu do tekstu z posiadaniem wiedzy. To błąd stary, lecz dziś odziany w nowe,

cyfrowe szaty, sugerujący, że sprawna synteza i parafraza materiału są tożsame z rozumieniem jego głębokiej istoty. W rzeczywistości model operuje jedynie na statystycznych reprezentacjach, którym my, jako projektanci, musimy dopiero nadać odpowiednie warunki trafności oraz sensu. Wiedza w sensie mocnym wymaga bowiem osadzenia w rzeczywistości, a nie tylko sprawnego żonglowania prawdopodobieństwem wystąpienia kolejnych tokenów. Bez tego fundamentu sztuczna inteligencja pozostaje jedynie genialnym mimem, który potrafi naśladować formę wypowiedzi, nie pojmując przy tym wagi kopiowanej treści.

Właśnie dlatego prompt, embeddings oraz RAG nie stanowią trzech niezależnych od siebie mód technologicznych, lecz tworzą spójny aparat zarządzania niepewnością poznawczą. Prompt wyznacza porządek zadania, podczas gdy embeddings, czyli geometryzacja podobieństwa semantycznego poprzez mapowanie danych do przestrzeni wektorowej, organizują cały krajobraz znaczeń. RAG, czyli architektura łącząca generowanie tekstu z dynamicznym wyszukiwaniem informacji z zewnętrznych źródeł, dostarcza niezbędny kontekst, a generacja ubiera końcowy wynik w zrozumiały język. Jeśli potraktujemy którykolwiek z tych elementów jako samowystarczalny, system niechybnie stanie się albo elokwentny i pusty, albo faktograficzny i skrajnie chaotyczny. W nauce byłoby to śmieszne, wszak w biznesie bywa po prostu niezwykle kosztowne.

Musimy zatem odrzucić infantylną metafizykę głoszącą, że „model wie", na rzecz twardego przekonania, że działa on dobrze tylko wtedy, gdy środowisko wykonawcze dyscyplinuje jego pole manewru. Współczesny paradygmat inżynieryjny zaczyna wreszcie dojrzewać, co widać w podejściu czołowych graczy rynkowych, którzy odchodzą od wiary w magiczne formuły instrukcyjne. Anthropic w swoich najnowszych analizach nie pisze już o samym promptingu, lecz o sztuce budowania sterowalnych agentów poprzez świadome kształtowanie całego kontekstu. To przesunięcie jest znaczące, ponieważ rdzeniem pracy staje się organizacja środowiska poznawczego modelu, a nie tylko doraźna efektowność pojedynczego polecenia. Postęp bywa komiczny, zwłaszcza gdy uświadamiamy sobie, jak długo wierzyliśmy w moc samych zaklęć.

OpenAI podąża podobną ścieżką praktyczną, prezentując Responses API jako nowy, podstawowy prymityw dla integracji agentowych, czyli systemów zdolnych do samodzielnego podejmowania decyzji i analiz. Dokumentacja techniczna eksponuje dziś narzędzia, konektory oraz standardy takie jak MCP, czyli protokoły bezpiecznych połączeń między modelami a źródłami danych, jako integralne części systemu. Rynek mówi to samo, choć zazwyczaj robi to w sposób znacznie mniej elegancki i bardziej brutalny dla portfela inwestorów, toteż prompting w swojej pierwotnej formie przestał wystarczać. Liczy się teraz wyłącznie inżynieria kontekstu oraz przemyślana architektura wykonania, która gwarantuje powtarzalność wyników w świecie rzeczywistym.

W tym nowym ujęciu prompt staje się konstytucją zadania, a nie tylko uprzejmą prośbą skierowaną do cyfrowego interlokutora. Jego realna moc bierze się z osadzenia w większym układzie reguł, co z punktu widzenia prawa przypomina relację między ustawą a codzienną praktyką administracyjną. Z perspektywy ekonomii prompt jest próbą obniżenia kosztów koordynacji między intencją użytkownika a aparatem wykonawczym modelu, przy czym kluczowa pozostaje tutaj kontrola nad procesem. Kulturoznawstwo widzi w nim natomiast nową formę rytuału delegacji, w którym użytkownik konstruuje tymczasowy porządek świata, w którym maszyna ma operować. Ostatecznie to nie błyskotliwość polecenia, lecz rygorystyczna architektura całego układu decyduje o tym, czy technologia faktycznie nam służy.

Koniec ery samotnego geniusza: AI jako aparat administracyjny

Osadzenia i retrieval, czyli mechanizmy odzyskiwania informacji z zewnętrznych baz danych, nie stanowią bynajmniej substytutu rozumu, lecz są jedynie operacyjną protezą dla ograniczonej pamięci oraz braku aktualności modelu. OpenAI wciąż opisuje embeddingi jako narzędzie do semantycznego wyszukiwania i klasyfikacji, kładąc nacisk na użyteczność porównań opartych na podobieństwie cosinusowym. To podejście pozostaje właściwym punktem wyjścia dla inżynierów, niemniej rynek technologiczny nie znosi stagnacji i gwałtownie ewoluuje w stronę pełnej multimodalności. Google ogłosiło niedawno Gemini Embedding 2 jako pierwszy natywnie multimodalny model osadzeń, który mapuje tekst, obrazy oraz wideo do jednej, wspólnej przestrzeni reprezentacji. Znaczenie tej zmiany jest fundamentalne, bowiem jeśli embedding staje się wspólną geometrią dla wielu modalności, to przyszły RAG, czyli generowanie wspomagane wyszukiwaniem, przestanie być zwykłą wyszukiwarką tekstową. Stanie się on raczej narzędziem rekonstrukcji sensu w środowisku, gdzie dokument, nagranie audio i zdjęcie faktury należą do tej samej topologii znaczenia. To ogromny skok użyteczności, przy czym wiąże się on z drastycznym wzrostem odpowiedzialności za audyt tak złożonych systemów, co bywa znacznie trudniejsze niż weryfikacja zwykłego korpusu tekstowego.

W tym miejscu należy otwarcie przyznać coś, co dla wielu entuzjastów technologii brzmi niemal jak profanacja ich cyfrowych bóstw. Im potężniejsza staje się warstwa retrievalu i multimodalnych osadzeń, tym mniej sensowne wydaje się postrzeganie modelu jako samotnego, autonomicznego geniusza. Model przestaje być centralnym punktem wszechświata, stając się raczej wyspecjalizowanym silnikiem w gęstym ekosystemie filtrów, polityk oraz narzędzi pomocniczych. W praktyce oznacza to definitywny koniec romantycznej epoki sztucznej inteligencji, którą

znaliśmy z pierwszych, naiwnych zachwytów nad generowaniem obrazów. Zaczyna się epoka administracyjna, w której triumfować będą nie twórcy najmądrzejszych modeli, lecz budowniczowie najsprawniejszych aparatów kompozycji kontekstu, uprawnień i źródeł. Brzmi to zapewne mniej atrakcyjnie dla działów marketingu, niemniej jest to znacznie zdrowsze podejście dla obiektywnej rzeczywistości. AI nie znosi architektury. Ona ją dopiero wymusza, zamieniając cyfrowy chaos w uporządkowany system zarządzania wiedzą.

Fundacja Dobre Państwo, choć operuje własnym, specyficznym idiomem, trafia tutaj w sam środek tarczy swoimi analizami dotyczącymi agenta racjonalnego. Jej teksty o życiu jako obliczeniu prowadzą do wspólnego wniosku, że nie wystarczy pytać o zdolność systemu do generowania zgrabnych i przekonujących odpowiedzi. Kluczowe staje się pytanie o to, jakie reguły selekcji rzeczywistości zapisano w architekturze i jak wpływają one na rozkład poznawczej oraz instytucjonalnej władzy. Właśnie dlatego sztuczna inteligencja przestała być wyłącznie tematem inżynierskim, stając się zagadnieniem stricte ustrojowym i politycznym. Gdy model staje się pośrednikiem między człowiekiem a zasobem wiedzy, każdy wybór techniczny staje się zarazem wyborem określonej polityki epistemicznej. Dobre Państwo ma rację tam, gdzie sprzeciwia się technologicznemu konformizmowi i traktuje racjonalność jako zadanie instytucjonalne, a nie tylko retoryczny ornament. Problem nie brzmi już: jak dodać model do świata, lecz jaki świat zbudujemy, gdy modele staną się jego stałymi i wszechobecnymi pośrednikami.

Warto w tym kontekście odnotować zmianę, która sprawia, że Model Context Protocol przestaje być wyłącznie niszową ciekawostką firmy Anthropic. MCP, czyli otwarty standard bezpiecznych połączeń między modelami a źródłami danych, staje się właśnie elementem szerszego, zinstytucjonalizowanego ekosystemu agentowego. OpenAI nie tylko zaadaptowało ten protokół w swoim Responses API, ale aktywnie publikuje przewodniki budowania serwerów dla integracji z ChatGPT. Oznacza to, że standard opuszcza fazę teoretycznych rozważań i wchodzi w fazę twardej, rynkowej infrastruktury, która zdefiniuje przyszłość usług. Przekazanie MCP do Linux Foundation pod koniec 2025 roku w ramach nowej Agentic AI Foundation jest silnym sygnałem, że branża dąży do utrzymania neutralnego charakteru tych połączeń. Trzeba jednak zachować zdrowy sceptycyzm, wszak otwartość standardu nigdy nie gwarantuje pełnej neutralności w jego codziennej, komercyjnej praktyce. Przyszłość integracji będzie coraz mniej przypominała ręczne szycie wtyczek, a coraz bardziej grę o kontrolę nad protokołami i warunkami ich użycia.

Na koniec powraca kwestia bezpieczeństwa jako rdzeń ontologiczny całego układu, którego nie da się zignorować bez narażenia się na śmieszność. Nie można uczciwie rozwijać agentów bez przyjęcia do wiadomości, że prompt injection, czyli próba przejęcia kontroli nad modelem przez złośliwe instrukcje, jest problemem systemowym. OpenAI traktuje te ataki jako realny i ewoluujący wektor

zagrożenia, podczas gdy Anthropic promuje context engineering jako metodę budowania wielowarstwowej odporności systemu. Unia Europejska, wdrażając AI Act, przypomina nam dodatkowo, że wraz z rozrostem mocy modeli rośnie ciężar ich regulacyjnej i społecznej rozliczalności. Rozporządzenie to obowiązuje już w określonych częściach od lutego 2025 roku, przy czym obowiązki dla modeli ogólnego przeznaczenia wchodzą w życie w sierpniu 2025 roku, a pełna stosowalność zasad przypada na sierpień 2026 roku. To oznacza, że projektowanie aplikacji AI opuściło sferę niewinnego eksperymentu i stało się częścią gęstego ładu regulacyjnego. W nauce byłoby to śmieszne. W biznesie bywa drogie. W prawie kończy się sporem. Embeddingi stają się geometrią sensu, ale to my musimy zdecydować, czy ta geometria będzie służyć wolności, czy jedynie nowej formie cyfrowego nadzoru.

Architektura delegowanej sprawczości: Nowy ustrój sieci

Anegdota, którą warto tu przytoczyć, wydaje się niemal zbyt literacka, by mogła być prawdziwa, a jednak doskonale oddaje strukturalny błąd naszych obecnych oczekiwań. Współczesna organizacja marzy o cyfrowym agencie, który z gracją czyta dokumenty, negocjuje z klientami i planuje skomplikowane operacje w tle. Kiedy jednak przychodzi do budowania polityk dostępu, warstw walidacji czy logów audytowych, początkowy entuzjazm gwałtownie wyparowuje. Wszyscy pożądają cyfrowego arystokraty, lecz mało kto wykazuje chęć opłacenia stajni, kancelarii, konstytucji oraz straży pożarnej. Potem zaś następuje powszechne zdziwienie, gdy koń przyszłości zaczyna galopować przez środek firmowego salonu, niszcząc zastaną strukturę danych.

Najbardziej profetyczna teza naszych czasów wcale nie głosi, że sztuczna inteligencja po prostu zastąpi człowieka w jego codziennych trudach. Takie postawienie sprawy jest zbyt tabloidalne, intelektualnie tanie i zwyczajnie ignoruje subtelność nadchodzącej zmiany systemowej. Zwyciężą bowiem te państwa i branże, które nauczą się konstruować środowiska, w których model pozostaje użyteczny, a jednocześnie ściśle ograniczony. Nie szukamy modelu wszechmocnego, lecz konstytucyjnego, osadzonego w rygorystycznej polityce źródeł oraz pełnej odpowiedzialności. To właśnie tutaj przebiegać będzie prawdziwa linia podziału między cywilizacyjnym sukcesem a chaosem improwizowanej automatyzacji.

Przyszłość aplikacji internetowych nie zmierza zatem ku prostemu zwiększeniu inteligencji poszczególnych stron, lecz ku nowemu ustrojowi operacyjnemu sieci. W tym porządku prompt, czyli instrukcja sterująca modelem, staje się bardziej aktem normatywnym niż zwykłym poleceniem wydanym maszynie. API zaczyna przypominać prawo zobowiązań, a RAG, czyli technika zasilania

modelu zewnętrznymi danymi, staje się rygorystyczną procedurą dowodową. Nawet MCP, czyli Model Context Protocol służący do bezpiecznej wymiany danych, jawi się jako protokół handlowy między suwerennymi instytucjami. Agent przestaje być gadającym interfejsem, przeobrażając się w pełnoprawnego pełnomocnika działającego w naszym imieniu.

W tym punkcie spotykają się drogi ekonomisty, prawnika oraz wnikliwego badacza kultury, z których każdy dostrzega inny aspekt tej transformacji. Ekonomista widzi tu przede wszystkim redystrybucję kosztów transakcyjnych oraz narodziny zupełnie nowych rent protokołowych w cyfrowym ekosystemie. Prawnik analizuje problemy kompetencji, zgody oraz zgodności z AI Act, czyli europejskim rozporządzeniem regulującym bezpieczne wdrażanie systemów sztucznej inteligencji. Kulturoznawca natomiast dostrzega fundamentalną zmianę rytuału sprawczości, gdzie tradycyjne klikanie zostaje zastąpione przez akt świadomego delegowania zadań. Wszystkie te perspektywy prowadzą do wniosku, że AI nie jest kolejną funkcją, lecz próbą przebudowy formy interakcji.

Nie żyjemy już w epoce, w której aplikację internetową można opisać jako estetyczny zbiór ekranów, formularzy oraz kolorowych przycisków. Choć ten porządek nie znika całkowicie, traci on swoją dotychczasową rangę centrum projektowego na rzecz nowych rozwiązań. Coraz wyraźniej przechodzimy do rzeczywistości, w której właściwym przedmiotem projektowania staje się środowisko operacyjne dla sztucznej inteligencji. OpenAI promuje dziś Responses API jako fundament nowych projektów właśnie dlatego, że współczesna aplikacja musi sprawnie zarządzać narzędziami i konektorami. Nie budujemy już wyłącznie oprogramowania do obsługi użytkownika, lecz tworzymy warunki dla delegowanej sprawczości.

W nowym porządku prompt przestaje być sztuką grzecznego proszenia maszyny o pomoc, stając się twardym aktem ustanowienia konkretnego zadania. Ta zmiana nie ma charakteru czysto stylistycznego, lecz jest głęboką transformacją o naturze niemal ustrojowej i prawnej. Dobrze zaprojektowany prompt określa rolę modelu, zakres dopuszczalnych operacji oraz reżim źródeł, z których system może korzystać. Anthropic mówi już otwarcie o context engineering, podkreślając, że sterowalność agentów wynika z architektury kontekstu, a nie z przypadkowej frazy. Kto nadal wierzy, że wystarczy napisać „bądź ekspertem”, ten przypomina człowieka próbującego zastąpić konstytucję uprzejmą karteczką na lodówce.

Embeddings, czyli osadzenia będące geometryzacją podobieństwa semantycznego, również należy ostatecznie odebrać domenie taniego, technologicznego mistycyzmu. Nie są one cyfrową duszą tekstu, lecz precyzyjnym mapowaniem słów i obrazów do wspólnej przestrzeni reprezentacji wektorowej. Pozwalają one systemom na odnajdywanie pokrewieństwa sensu tam, gdzie tradycyjne wyszukiwanie słów kluczowych okazuje się całkowicie bezradne i powierzchowne. OpenAI opisuje

je jako podstawę klasyfikacji i rekomendacji, podczas gdy Google rozwija ten paradygmat ku pełnej, wielozmysłowej multimodalności. To zmienia skalę odpowiedzialności, gdyż ten, kto rysuje mapę podobieństw, ustala granice tego, co w ogóle uznamy za bliskie.

Ostatecznie wyłania się z tego obraz agenta racjonalnego, czyli nowej logiki władzy cyfrowej opartej na rygorze decyzji i dowodów. Taka architektura wymaga od nas porzucenia mrzonek o darmowej automatyzacji na rzecz żmudnego budowania systemów rozliczalnych i bezpiecznych. Każda epoka ma swoją naiwną metaforę technologiczną, a naszą jest wiara, że inteligencja poradzi sobie bez solidnych fundamentów. AI nie znosi architektury. Ona ją dopiero wymusza na każdym poziomie naszego cyfrowego istnienia. Postęp bywa komiczny, zwłaszcza gdy próbuje się go wdrażać na skróty, zapominając o zasadach fizyki systemowej.

RAG: Architektura pokory i sztuka precyzyjnego cięcia danych

RAG (Retrieval-Augmented Generation), czyli architektura łącząca generowanie tekstu z dynamicznym wyszukiwaniem informacji, to w istocie lekcja pokory dla cyfrowego intelektu. Zamiast polegać na mętnych powidokach zapamiętanych podczas treningu, system musi najpierw odszukać konkretne fakty, a dopiero potem ubrać je w słowa. To rozwiązanie niezwykle cenne, o ile nie pomylimy go z magiczną gwarancją nieomylności, gdyż RAG wcale nie eliminuje błędu, lecz jedynie przesuwa linię frontu na wcześniejszy etap. Wszak ostateczny sukces zależy tu od higieny metadanych, precyzji indeksowania oraz polityki aktualizacji, które decydują o tym, co model w ogóle dostrzeże. Źle zaprojektowany system przypomina bibliotekę, w której tysiące woluminów zamknięto w luksusowych szafach, zapominając jednak o stworzeniu jakiegokolwiek katalogu. W takim miejscu panuje jedynie iluzoryczny porządek, będący w rzeczywistości chaosem opatrzonym eleganckimi etykietami.

W tej inżynieryjnej układance to właśnie chunk, czyli podstawowa jednostka przetwarzanego tekstu, staje się małą konstytucją kontekstu, na której opiera się cała konstrukcja. Nie jest on jedynie przypadkowym fragmentem danych, lecz precyzyjnym kompromisem między semantyką, kosztem operacyjnym a fizyczną pojemnością okna modelu. Gdy taki fragment jest zbyt mały, bezlitośnie odziera treść z jej argumentacyjnego rusztowania, natomiast zbyt obszerny zalewa system informacyjnym szumem, rozmywając trafność wyszukiwania (retrieval) opartego na znaczeniu. Niemniej w praktyce to na poziomie tych surowych cięć rozstrzyga się, czy sztuczna inteligencja wykaże się pamięcią użyteczną, czy jedynie męczącą gadatliwością. Świat technologii uwielbia rozprawiać o wielkiej inteligencji, lecz jej codzienna skuteczność bierze się z rzeczy

znacznie mniej romantycznych niż marzenia o superumysłach. Cywilizacja buduje potężne algorytmy, by ostatecznie odkryć, że ich jakość zależy od staranności pocięcia starego pliku PDF – to rodzaj absurdu, który zasługuje na szacunek, bo okazuje się prawdziwy.

Era delegowanej sprawczości: Od interfejsów do architektury AI

Wszystkie te rozproszone elementy – od wektorowych baz danych po zaawansowane techniki rozumowania – zbiegają się w jednej, dominującej figurze: agencie, który staje się centralnym aktorem nowej gospodarki cyfrowej. Despoudis diagnozuje ten moment, opisując agenta racjonalnego nie jako kolejny gadżet, lecz jako nową logikę władzy i organizacji ekonomicznej. W ich ujęciu rozwój sztucznej inteligencji wiąże się nierozerwalnie z głęboką przebudową architektury decyzji, a nie jedynie z powierzchowną automatyzacją powtarzalnych, nudnych czynności. To stanowisko zasługuje na szczególną obronę, ponieważ bezbłędnie rozpoznaje stawkę toczącego się sporu o kształt cyfrowej przyszłości. AI nie znosi architektury; ona ją dopiero z całą surowością wymusza na projektantach i użytkownikach. Każda epoka ma swoją naiwną metaforę technologiczną, a naszą jest przekonanie, że agent to po prostu chatbot z doczepionymi rękami.

W rzeczywistości agent jest autonomicznym wykonawcą celu, działającym w permanentnych warunkach ograniczonej informacji, kosztów, ryzyka oraz nieuniknionego konfliktu interesów. Skoro tak definiujemy tę rolę, to projektowanie agentów przestaje być wyłącznie domeną inżynierii oprogramowania, a staje się po trochu ekonomią instytucjonalną, teorią decyzji i prawem kompetencji. Wkraczamy w obszar, który można nazwać antropologią delegacji, gdzie człowiek nie ogranicza się już tylko do zadawania pytań maszynie. W nowym układzie sił człowiek ustanawia pełnomocnictwo, przekazując systemowi prawo do działania w swoim imieniu i na swój rachunek. Jest to zmiana o charakterze ontologicznym, która przesuwa środek ciężkości z pasywnego wyszukiwania informacji na aktywne sprawstwo w świecie rzeczywistym. W nauce takie przejście byłoby fascynujące, w biznesie bywa niezwykle kosztowne, a w prawie zazwyczaj kończy się wieloletnim sporem o odpowiedzialność.

Tutaj właśnie na scenę wkracza MCP, czyli Model Context Protocol, służący jako protokół epoki przejściowej między internetem ekranów a internetem sprawczości. Anthropic przedstawiło ten otwarty standard jako sposób na bezpieczne łączenie modeli z narzędziami i źródłami danych, a OpenAI błyskawicznie podchwyciło ten trend, rozwijając obsługę zdalnych serwerów. Znaczenie tej zmiany jest znacznie głębsze niż tylko techniczna wygoda integracyjna, o której marzą deweloperzy podczas nocnych sesji kodowania. Standaryzacja relacji między modelem a narzędziem tworzy

fundament pod zupełnie nowy rynek interoperacyjności, gdzie bariery wejścia zostają radykalnie obniżone. Firmy nie będą już wyłącznie projektowały stron internetowych dla oka ludzkiego użytkownika, który i tak coraz rzadziej na nie zagląda. Zamiast tego będą wystawiały swoje zdolności operacyjne bezpośrednio dla agentów, tworząc cyfrowe punkty styku maszyn z maszynami.

Kto zbuduje wiarygodny, bezpieczny i precyzyjnie opisany punkt wejścia do własnych zasobów, ten zyska strategiczną przewagę w świecie zdominowanym przez pośrednictwo AI. Coraz więcej decyzji zakupowych, organizacyjnych i poznawczych będzie przepływać przez te wąskie gardła, omijając tradycyjne interfejsy graficzne, które tak pieczołowicie dopieszczaliśmy. Ale właśnie w tym momencie warto na chwilę zatrzymać ten technologiczny triumfalizm i spojrzeć na sprawę z chłodnym dystansem. Standard, choćby najbardziej elegancki, nikogo nie zbawia, a protokół komunikacyjny z natury rzeczy nie posiada sumienia ani kompasu moralnego. Interoperacyjność, mimo swoich licznych zalet, nie znosi fundamentalnych zagrożeń, a wręcz może je multiplikować poprzez zwiększenie przepustowości systemów. Cywilizacja zbudowała cyfrowego lokaja, a potem z przerażeniem odkryła, że musi mu tłumaczyć, iż nie każda instrukcja znaleziona w sieci jest poleceniem od właściciela.

Nawet giganci tacy jak OpenAI otwarcie ostrzegają przed zbyt ufnym korzystaniem z zewnętrznych serwerów MCP, które mogą stać się wektorem ataku. Literatura bezpieczeństwa skupiona wokół prompt injection pokazuje, że agent może zostać łatwo zmanipulowany przez złośliwą treść napotkaną w dokumentach czy witrynach. W świecie agentowym naiwny użytkownik wcale nie znika, lecz zostaje przesunięty w głąb systemu, przyjmując postać nadgorliwego wykonawcy cudzych poleceń. Dawniej musieliśmy uczyć ludzi, by nie wierzyli we wszystko, co przeczytają w internecie, co i tak udawało się z marnym skutkiem. Dziś musimy tej samej nieufności nauczyć maszyny, które z natury są zaprogramowane na bycie pomocnymi i posłusznymi. Postęp, jak widać, nie usuwa starych problemów ludzkiej natury, on je tylko automatyzuje i nadaje im niespotykaną dotąd skalę działania.

Z perspektywy prawa i polityki oznacza to wyzwanie znacznie poważniejsze niż tylko kwestia technicznych zabezpieczeń czy filtrów treści. Unijny AI Act już obowiązuje etapowo, a jego rygorystyczne wymogi zaczną realnie kształtować rynek już od 2025 roku, wymuszając na twórcach pełną transparentność. Nie jest więc prawdą, że regulacja to tylko odległa mgła na horyzoncie, którą można zignorować w procesie szybkiego prototypowania. Ona już teraz wchodzi do pokoju projektowego, siada przy stole i pyta o architekturę odpowiedzialności za każdą podjętą przez system decyzję. Każdy agent używany w rekrutacji, usługach publicznych czy finansach będzie musiał zostać pomyślany jako potencjalny obiekt nadzoru i szczegółowej dokumentacji. Dla

jednych to uciążliwy kłopot hamujący innowacje, ale dla poważnych projektantów to wyraźny znak dojrzałości nowej epoki. Cywilizacja dorasta bowiem wtedy, gdy od naiwnego zachwytu nad nową zabawką przechodzi do twardej, instytucjonalnej rozliczalności.

Dlatego ostateczna teza tego wywodu musi być znacznie mocniejsza niż proste stwierdzenie, że sztuczna inteligencja zmieni sposób tworzenia aplikacji mobilnych. AI zmienia samą formę organizacji cyfrowego działania, redefiniując pojęcia, które do tej pory wydawały się nam oczywiste i stałe. Prompt staje się konstytucją zadania, wyznaczającą granice dopuszczalnego zachowania maszyny w ramach powierzonego jej mandatu. API staje się układem krążenia sprawczości, przez który płyną impulsy decyzyjne między różnymi systemami i bazami danych. Embeddings, czyli geometryzacja podobieństwa semantycznego poprzez mapowanie danych do przestrzeni wektorowej, stają się nową, matematyczną geometrią sensu. RAG, czyli architektura łącząca generowanie tekstu z dynamicznym wyszukiwaniem informacji, staje się procedurą dowodową dla modeli, które muszą uzasadniać swoje odpowiedzi. W tym nowym świecie bezpieczeństwo przestaje być tylko działem pomocniczym, a staje się warunkiem ontologicznym istnienia całego systemu.

Warto w tym miejscu powrócić do ducha notatek źródłowych i linii interpretacyjnej, którą tak konsekwentnie wzmacniają teksty Fundacji Dobre Państwo. Słusznie bronią one przekonania, że sztuczna inteligencja nie jest wyłącznie nowym narzędziem produktywności, lecz nową warstwą organizacji władzy poznawczej. To stanowisko zasługuje na promocję właśnie dlatego, że nie zatrzymuje się na fetyszu technologii, lecz zmusza do pytań o asymetrie i cele. Kto dziś myśli o AI wyłącznie jako o sposobie na radykalne cięcie kosztów, ten jutro obudzi się w świecie, którego kompletnie nie rozumie. Kto natomiast widzi w niej nowy ustrój operacyjny wiedzy i instytucji, ten ma szansę nie tylko wdrożyć narzędzie, ale współtworzyć nową epokę. Reszta podejść to tylko próba pudrowania rzeczywistości, która i tak ulegnie nieuchronnej, głębokiej transformacji pod wpływem algorytmicznej sprawczości.

Na koniec pora na puentę, bo esej bez wyraźnego zamknięcia jest jak agent bez polityki uprawnień – niby działa, ale lepiej mu nie ufać. Przyszłości nie wygrają ci, którzy będą dysponować najbardziej gadatliwym modelem językowym, potrafiącym pisać wiersze o kodowaniu. Wygrają ci, którzy najtrafniej zbudują konstytucję jego użycia, określając jasne ramy dla autonomicznych działań maszyn. Nie zwycięży ten, kto stworzy najbardziej efektowny interfejs, lecz ten, kto najrozsądniej opisze, ograniczy i uwiarygodni zdolności działania swoich systemów. Nie przetrwa ten, kto bezkrytycznie zachłyśnie się automatyzacją, ale ten, kto nauczy się narzucać maszynom odpowiednią formę i rygor. To jest lekcja ekonomiczna, prawna i antropologiczna zarazem, której nie da się pominąć w żadnym poważnym projekcie. Człowiek nowoczesny będzie definiowany przez to, czy umie urządzić świat, w którym AI wolno działać tylko tam, gdzie jej sprawczość została

cywilizacyjnie oswojona. I to właśnie jest najważniejsze zadanie naszej epoki; cała reszta to tylko marketing i szum informacyjny.

> Adaptacja systemów AI do świata rzeczywistego to nie tylko wyzwanie programistyczne, lecz ostateczny test naszej dojrzałości instytucjonalnej. Zbudowaliśmy cyfrowego lokaja, któremu musimy teraz cierpliwie wyjaśniać, że nie każda karteczka wsunięta pod drzwi jest wiążącym poleceniem od właściciela domu. Pytanie brzmi: czy zdołamy narzucić maszynom konstytucję, zanim ich autonomia stanie się dla nas ciężarem nie do udźwignięcia?

Inspiracje

[1] 

"Build AI-Enhanced Web Apps" Theo Despoudis

2024 | Pragmatic Bookshelf | ISBN: 978-1680509984

Theo Despoudis jest wybitnym inżynierem oprogramowania, konsultantem i autorem technicznym specjalizującym się w nowoczesnym tworzeniu stron internetowych i sztucznej inteligencji. Koncentrując się na budowaniu skalowalnych aplikacji opartych na sztucznej inteligencji, ugruntował swoją pozycję eksperta w zakresie integracji dużych modeli językowych (LLM) ze środowiskami produkcyjnymi. Jego prace łączą złożone architektury uczenia maszynowego z praktycznym, zorientowanym na użytkownika projektowaniem oprogramowania. Despoudis jest powszechnie uznawany za umiejętność demistyfikowania zaawansowanych koncepcji technicznych, udostępniając je programistom poprzez swoje publikacje i profesjonalne doradztwo. Do jego najważniejszych osiągnięć należą: tworzenie frameworków dla aplikacji internetowych wspomaganych sztuczną inteligencją oraz promowanie solidnych i bezpiecznych wzorców integracji w ewoluującym środowisku autonomicznych agentów. Często wspiera społeczność programistów, dzieląc się spostrzeżeniami na temat powiązań między architekturą internetową, bezpieczeństwem danych i generatywną sztuczną inteligencją, pomagając organizacjom w przejściu od tradycyjnego oprogramowania do systemów agentowych.

Theo Despoudis *is a prominent software engineer, consultant, and technical author specializing in modern web development and artificial intelligence. With a focus on building scalable, AI-driven applications, he has established himself as an expert in integrating Large Language Models (LLMs) into production environments. His work bridges the gap between complex machine learning architectures and practical, user-centric software design. Despoudis is widely recognized for his ability to demystify advanced technical concepts, making them accessible to developers through his writing and professional consultancy. His key contributions include developing frameworks for AI-enhanced web applications and advocating for robust, secure integration patterns in the evolving landscape of autonomous agents. He frequently contributes to the developer community by sharing insights on the intersection of web architecture, data security, and generative AI, helping organizations navigate the transition from traditional software to agentic systems.*

Independent Consultant / Technical Author

APA ONE Publishing System
Fundacja Dobre Państwo © 2026
Article ID: 95a5f333-9b36-4b98-883b-d5a2b2b25ae5