AI na progu 2026 – od fascynacji do dojrzałości?

Wyzwania w drodze do lepszego AI

Remigiusz Kinas *

remigiusz.kinas@gmail.com

Principal AI Researcher w Ingenix.ai, współautor polskiego modelu językowego Bielik.

19 stycznia 2026

Streszczenie

Artykuł syntetyzuje stan rozwoju AI na progu 2026 roku i argumentuje, że główny kierunek postępu przesuwa się z ilościowego skalowania modeli ku jakościowej dojrzałości systemowej. Analizuję trzynaście obszarów badawczych (m.in. continual learning, optymalizacja kompresji wiedzy, test-time compute, modele świata, agentowość i multimodalność) oraz ich konsekwencje dla niezawodności i bezpieczeństwa AI. Teza zasadnicza brzmi czy kluczowym wyzwaniem nadchodzących lat jest dalszy wzrost sprawności poznawczej, czy domknięcie luki między rosnącą sprawczością modeli a możliwościami ich interpretacji, diagnozy i kontrolowalności. Pokazuję, że granicą „lepszego AI" stają się nie tylko dane i architektury ale również koszty energii oraz infrastruktura obliczeniowa, które nadają rozwojowi wymiar geopolityczny. Esej zamykam dyskusją o granicach antropomorfizacji i o tym jak zmienia się ludzka podmiotowość w świecie optymalizowanym przez algorytmy o nadludzkiej wydajności wnioskowania.

Słowa kluczowe: AGI, continual learning, metamyślenie, modele świata, interpretowalność AI, scaling laws, agentowość, bezpieczeństwo AI, Open Source.

*W artykule tym AI pomagało w zakresie: tłumaczenie zawiłych tekstów i pojęć na j. polski (DeepL). Oryginał powstał w j. polskim, a przetłumaczył go na język angeilski Claude Code wspomagany modelem Opus 4.5 (ręczna korekta). AI użyte było również do sprawdzania literówek i błędów w pisowni. Gemini-3-Pro-Flash do redakcji niektórych zdań w zakresie stylistyki. Overleaf z wtyczką do LLM jako edytor LaTeX – wskazówki poprawy stylistyki tekstu. Research za pomocą wyszukiwarek powered by AI. Do zrozumienia, na przestrzeni całego roku, publikacji źródłowych wykorzystano OpenAI ChatGPT od 4 do 5.2 oraz rodzinę modeli Gemini 3. Szablon blogu powstał za pomocą Claude Code (vide coding i ręczna korekta). Wszystkie pomysły na kategoryzację, przemyślenia oraz rozwinięcie tematu wymyślone przez człowieka i napisane ludzką ręką.

Spis treści

1 Wstęp

Dotychczasowy paradygmat oparty na wykładniczym zwiększaniu mocy obliczeniowej i ilości danych treningowych zderza się z nowymi barierami zarówno technologicznymi, jak i fizycznymi. Przechodzimy z etapu fascynacji możliwościami generatywnymi (choć jestem pewien, że w tym roku zostaniemy wielokrotnie zaszokowani umiejętnościami modeli GenAI, czy nowym otwarciem w robotyce) do fazy, w której kluczowa staje się niezawodność, efektywność energetyczna oraz zdolność systemów do ciągłej adaptacji. Przechodzimy z ery skalowania w erę badań i optymalizacji oraz zwiększania efektywności nie przez rewolucyjne działania, a przez efekt ciągłego doskonalenia.

Należy jednak uczciwie odnotować, że teza o wyczerpaniu się paradygmatu „brutalnego" skalowania (brute-force scaling) pozostaje w 2026 roku przedmiotem ostrego sporu. Głosy takie jak Leopolda Aschenbrennera [1] przekonują, że scaling laws wcale nie wyhamowały, a jedynie zmieniły „paliwo" – z surowych danych internetowych na gigantyczne ilości danych syntetycznych oraz potężne zasoby obliczeniowe poświęcone na samo wnioskowanie (test-time compute). Z tej perspektywy droga do AGI nie wiedzie przez algorytmiczną elegancję, lecz przez budowę klastrów wartych setki miliardów dolarów, które samą masą krzemu i energii przesuwają granice inteligencji. Stoimy więc przed pytaniem: czy w wyścigu o prymat zwycięży system najbardziej „sprytny", czy ten, za którym stoi największa elektrownia atomowa?

Nie przypadkowo w tytule użyłem bezpiecznego dla mnie określenia „lepszego AI" ponieważ moim zdaniem rok 2026 zmusza nas do powrotu do definicji minimalnego AGI (Minimal AGI). Shane Legg z Google DeepMind [2] definiuje je jako moment, w którym system potrafi wykonać każde zadanie poznawcze, jakie jest w stanie wykonać człowiek. Nie szukamy już tylko geniusza rozwiązującego zagadki matematyczne, a szukamy „przeciętności", która jest uniwersalna. Nie pytamy już, czy AGI jest możliwe, ale na którym poziomie tego spektrum się znajdujemy. Przypominamy sobie bowiem „wpadki" wielkich graczy AI, którzy co prawda potrafili rozwiązywać złożone problemy matematyczne, jednocześnie potykając się na prostych zadaniach typu „strawberry" (choć nie jest to najlepsze zadanie dla LLM'a ale obnaża jego obecne AGI).

Ostatnie kilka tygodni i przejście w rok 2026 skłoniło mnie do refleksji nad przemijającymi dokonaniami AI oraz przyszłością czyli tym co może przynieść nam kolejny rok. Im dłużej pisałem ten artykuł, tym więcej pojawiało się w mojej głowie pytań, przemyśleń i wątpliwości. Czy to wyzwania obecnego roku, czy może roadmapa na kolejne kilka, a może kilkanaście lat? A przecież tyle mówi się obecnie o AGI. Najwięksi tego świata, firmy produkujące znane systemy AI licytują się w przewidywaniu daty nadejścia super inteligencji - za rok, za pięć, za dziesięć lub wcale (nie ma jednoznaczności w tym zakresie). Czy zatem wiedzą oni coś czego my zwykli śmiertelincy, użytkownicy AI, nie widzimy? Czy wielkie amerykańskie i chińskie laby skrywają przed nami AGI - super inteligencję, która w przyszłości rozwiąże największe problemy tego świata? A może spodziewając się widma nachodzącej klęski realizacji AGI w

najbliższych latach zawęzili dla niej wymagania do rozwiązywania wąskiego rodzaju zadań np. bycia mądrym chatem na poziomie wykładowcy akademickiego (częsty slogan twórców AI - „model na poziomie doktora")? Później naszło mnie jeszcze ważniejsze pytanie - jak będzie wyglądała ludzka przyszłość za kilka, kilkanaście lat?

Nie mam jednak wątpliwości, że temat AGI jest ostatnio bardzo modny. Często nie jest jednak dogłębnie analizowany. Wyrażane są ogólne stwierdzenia, w których brak szerokiej analizy stanu obecnego AI. Prezentowane są opinie bez konkretnej definicji wymagań AGI, stanu docelowego. Definiuje się kolejne pojęcia ASI (Super Intelligence), które jeszcze bardziej zaciemniają obraz. Moim zdaniem ze szkodą dla AI, dla zrozumienia czym ono jest i dokąd zmierzamy. Rzadko mówi się też o wyzwaniach, problemach i kierunkach prac dzisiejszej sztucznej inteligencji. Widzimy jednak postęp, większoć z nas doświadcza AI i jest odbiorcą korzyści jakie daje ta technologia. Postęp ten z dnia na dzień przyspiesza. To już nie drobne kroczki a siedmiomilowe kroki. Mam osobiste wrażenie, że na początku 2026 roku każdy dzień przynosi większe zmiany niż tygodniowy postęp w 2024. Jednocześnie, zgodnie z tym co twierdzą twórcy LLM Arena, najlepsze modele pozostają na topie średnio 35 dni, spadają poza TOP5 w przeciągu pięciu miesięcy. Claude 3 Opus, wprowadzony w marcu 2024, plasuje się obecnie na 139 miejscu listy LLM Arena.

Wejście w rok 2026 przynosi twarde dowody na to, że bariera złożoności zadań, które stawiamy przed AI, systematycznie maleje. Wystarczy spojrzeć na benchmark *FrontierMath Tier 4* [3] uchodzący za bastion nierozwiązywalnych problemów matematycznych. Z 48 bardzo trudnych zadań, aż 14 uległo mocy flagowego modelu OpenAI, GPT-5.2 (Pro) (stan na 11.01.2026). Jestem wręcz pewien, że w tym roku padną kolejne zadania ze zbioru FrontierMath. Obstawiam, że przekroczymy ponad 50% rozwiązań. Z najnowszych osiągnięć, właściwie w trakcie pisania tego esseju, zaskakują nas kolejne informacje z obszaru matematyki badawczej. Na erdosproblems.com odnotowano przypadek, w którym model z rodziny GPT-5.2 doprowadził (w pętli z człowiekiem oraz formalizacją w Lean) do rozstrzygnięcia kilku problemów Erdősa. Co istotne, sukces nie wynikał z „magicznej intuicji" jednego promptu, tylko z procesu. Barierą nie była już sama „moc obliczeniowa", tylko kontrola halucynacji i domykanie luk w dowodzie przez iteracyjną krytykę oraz wsparcie procesem formalnego dowodzenia. Nawet jeśli część takich „rozwiązań" bywa później zredukowana do odnalezienia wyniku w literaturze albo doprecyzowania niejednoznacznej treści zadania, sam fakt, że pętla *LLM → poprawki → formalizacja* działa w praktyce, przesuwa granicę tego, co uznajemy za wykonalne przez AI w 2026 roku. Jesteśmy u progu nowej rewolucji przemysłowej. O ile poprzednia zastąpiła ludzkie mięśnie, obecna zastępuje funkcje mózgu.

Kolejną ewolucję obserwujemy w benchmarku *ARC-AGI-2* [4]. Rok 2025 rozpoczynaliśmy z wynikiem na granicy kilku procent, by zamknąć go z całkiem imponującą skutecznością na poziomie 54.2% (również model OpenAI, GPT-5.2 (Pro)) i propozycją nowego testu ARC-AGI-3 [5]. Porównywanie modeli na części znanych benchmarków GPQA Diamond, HMMT, AIME 2025, MMMLU przestaje mieć moc dyskryminacyjną (wyniki powyżej 90%). Testy te co najwyżej mogą służyć jako tzw. sanity check do weryfikacji jakości treningu. Mogą

potwierdzać, że model nie uległ degradacji (regresji) i nie „zapomniał" fundamentalnych zasad logiki w wyniku błędu inżynieryjnego.

A co z vibe? Rozwijają się vibe-coding, vibe-designing (muzyka, grafika). Pamiętam swoje pierwsze próby współpracy z agentami do kodowania. Zwykle ilość czasu poswięconego na poprawianie błędów agenta była większa niż pisanie od zera. Być może to moja nieudolność, brak wiedzy. Sytuacja zmieniła się jednak pod koniec roku. Kolejne rozwiązania wspaniale spełniają swoje zadanie. Może nie idealnie, ale zauważalnie lepiej. Pomagają nie tylko zawodowcom ale również tym, którzy nigdy w życiu nie podjęliby się programowania. Nawet ortodoksi kodujący „od zera" np. Linus Torvalds wspomina o realnych korzyściach z użycia AI do pisania kerneli Linux'a. To jednak nie koniec. Kiedy OpenAI definiuje poziomy dojrzałości AI stawiając jako cel poziom piąty *„Organizations: AI that can do the work of an organization"* pojawiają się pierwsze rozwiązania na vibe business np. Atoms.dev [6]. Środowisko programistyczne oparte na koncepcji „AI Team". Pozwala ono przekształcać naturalne opisy pomysłów w gotowe produkty cyfrowe. Atoms buduje „wirtualny" zespół deweloperski, w którym autonomiczne jednostki dzielą pracę na etapy planowania, projektowania architektury, pisania kodu full-stack oraz wdrażania i testowania. Wszystko praktycznie z minimalnym udziałem człowieka.

W niniejszym artykule zdefiniowałem trzynaście obszarów, które moim zdaniem zadecydują o dalszym postępie AI, ale też o tym jak będzie wyglądała przyszłość relacji człowiek-technologia a tym samym jak będzie wyglądał świat. Od „data wall" i konieczności wyjścia poza ludzkie dane, przez wyzwania związane z pamięcią i wnioskowaniem na metapoziomie, aż po budowanie podwalin przyszłej symbiozy człowieka z maszyną. To kilka kierunków do tego, by AI przestało być jedynie statycznym archiwum wiedzy a stało się dynamicznym, rozumującym systemem zdolnym do czegoś więcej niż sprytne odtwarzanie. W artykule powołuję się na najnowsze publikacje definiując nimi stan obecny zagadnień i jednocześnie punkt wejścia w rok 2026. Opisane punkty pozwolą mi systematycznie śledzić rozwój technologii w drodze do lepszego AI (AGI). Nie ma jednak róży bez kolców. Idealne AI i chęć zaadresowania większości tych punktów spowoduje, że AI będzie znacznie silniejsze od człwieka. Co zatem z bezpieczeństwem, interpretowalnością, przyszłym światem? To pewnie temat już na zupełnie inny artykuł. Miejmy to jednak na uwadze.

2 Ciągłe uczenie i adaptacja (continual learning)

Najbardziej zdradliwą cechą współczesnych modeli nie jest brak „inteligencji", tylko jej nierówny jakościowo charakter (jagged intelligence). Jak to rozumieć? Skrajna biegłość obok zaskakująco prymitywnych błędów. Prawdziwą metryką 2026, jak wspomina Demis Hassabis z DeepMind [7], będzie więc spójność i niezawodność, a nie kolejne punkty na benchmarkach. Jak pomóc AI w samodoskonaleniu (adaptacji, „w locie", do nowych zadań) przy jednoczesnym doświadczaniu wielu przeszkodód takich jak model drifting, catastrophic forgetting, contamination (jak ocenić nową informację), niewystarczająca moc obliczeniowa, skuteczność metod?
Jedną z takich opcji jest ciągłe uczenie [8] [9]. Continual learning (CL) to paradygmat odchodzący od statycznego modelu trenowania na zamrożonych zbiorach danych na rzecz systemów dynamicznych, które ewoluują wraz z napływającymi informacjami. W kontekście LLM (Large Language Models), wyzwanie to polega na efektywnej adaptacji do zmieniających się dystrybucji danych bez konieczności kosztownego trenowania modelu od zera za każdym razem, gdy pojawią się nowe fakty, regulacje prawne czy preferencje użytkowników.

Zgodnie z najnowszymi analizami, CL w świecie wielkich modeli językowych realizuje się w dwóch głównych kierunkach:

- **Ciągłość wertykalna (vertical continuity):** polega na stopniowym przechodzeniu od ogólnych zdolności modelu do wysoce specjalistycznych kompetencji. Proces ten obejmuje trzy etapy: ciągłe douczanie wstępne (Continual Pre-Training – CPT), adaptację domenową (Domain-Adaptive Pre-training – DAP) oraz ciągłe dostrajanie instrukcyjne (Continual Fine-Tuning – CFT).
- **Ciągłość horyzontalna (horizontal continuity):** skupia się na zdolności modelu do adaptacji w czasie i w poprzek różnych domen, pozwalając mu na przyswajanie nowych trendów i faktów przy jednoczesnym zachowaniu wiedzy historycznej.

Kluczową barierą technologiczną pozostaje tzw. katastrofalne zapominanie (catastrophic forgetting). Zjawisko to występuje, gdy model podczas nauki nowych informacji nadpisuje parametry odpowiedzialne za wcześniej nabyte umiejętności, co prowadzi do gwałtownego spadku wydajności w starych zadaniach. Rozwiązania CL mają na celu stworzenie mechanizmów „ukierunkowanej adaptacji", które są znacznie bardziej wydajne zasobowo pozwalają na aktualizację modelu przy ułamku kosztów obliczeniowych pełnego treningu.
Wyzwania, które stoją przed nami, a które są już po części adresowane w badaniach można podzielić na cztery kategorie:

A. **Architektura** – dynamika architektury – czy warunkiem do doskonałości i uzyskania AGI jest umiejętność zmiany struktury fizycznej modelu?
   Czy pozostaniemy przy MoE (Mixture of Experts) [10] pozwalające modelowi na dynamiczny wybór wyspecjalizowanych „ekspertów" (podsieci) dla każdego przetwarzanego

tokena. A może przy *Mixture-of-Depths* [11] zakładające, że model decyduje dynamicznie, które tokeny wymagają pełnego przetworzenia przez warstwy transformera, a które mogą je pominąć. Rozwiazanie takie pozwala na powiedzmy „inteligentne" alokowanie mocy obliczeniowej w czasie rzeczywistym. Oznacza to przejście od sztywnego przetwarzania w taki sam sposób każdego elementu danych do architektury, która uczy się selektywnie angażować swoje zasoby tylko tam, gdzie wymaga tego złożoność informacji. Czy w najbliższym czasie będziemy obserować bardziej radykalne odkrycia, ukierunkowane na częściową lub całkowitą zmianę architektury modelu „w locie"?

B. **Trening** – jak dynamicznie zmieniać umiejętności modelu po jego wdrożeniu wykorzystując techniki treningowe?
W 2025 roku pojawiają się pierwsze konkretne mechanizmy trwałej adaptacji modeli. Na przykład Self-Adapting Language Models (SEAL) [12]. W metodzie tej LLM generuje własne dane treningowe („self-edits") i wykorzystuje je do aktualizacji swoich wag za pomocą pętli reinforcement learning (RL). Dzięki temu model trwale uczy się nowych informacji bez konieczności treningu od zera. Krok w stronę LLM-ów, które faktycznie aktualizują się i adaptują na podstawie własnych doświadczeń (więcej o tym piszę w rozdziale "3. Doświadczanie ponad ludzkie dane").

   Innym podejściem jest BDH (Dragon Hatchling: The Missing Link between the Transformer and Models of the Brain) [13], które rezygnuje ze sztywnego podziału na fazę treningu i wnioskowania na rzecz mechanizmów hebbowskich (Hebbian Learning). Zamiast polegać wyłącznie na wstecznej propagacji błędu, system naśladuje biologiczną plastyczność. Neurony, które wspólnie reagują na dany bodziec, wzmacniają swoje połączenia w czasie rzeczywistym. Sprawia to, że nauka staje się naturalnym efektem ubocznym przetwarzania informacji, eliminując efekt „Dnia Świstaka", w którym model zapomina kontekst interakcji natychmiast po jej zakończeniu.

   Koniec 2025 roku przynosi jednak kolejne propozycje w postaci paradygmatu *Nested Learning* [14]. Zaprezentowany tam moduł *Hope* (co prawda na razie w fazie badawczej) to system „samomodyfikujący się" (self-modifying). Jego innowacyjność polega na zerwaniu z „iluzją sztywnej architektury", gdzie model (wagi) jest oddzielony od statycznego algorytmu uczenia (optymalizatory np. Adam, SGD itd.). W podejściu Nested Learning algorytm optymalizacji staje się częścią samej sieci. Moduł *Hope* działa w pętli zagnieżdżonej: podczas gdy warstwy bazowe przetwarzają dane, moduł nadrzędny analizuje dynamikę błędów i w czasie rzeczywistym przepisuje reguły aktualizacji wag dla konkretnych neuronów. Dzięki temu sieć może lokalnie zwiększać plastyczność dla nowych zadań a jednocześnie „zamrażać" regiony odpowiedzialne za starą wiedzę. Jednocześnie rozwiazanie to zmniejsza problem katastroficznego zapominania (catastrophic forgetting) na poziomie samej matematyki uczenia, a nie tylko architektury.

C. **Memory consolidation** – jak przenieść coś z pamięci krótkotrwałej (kontekst) do długotrwałej (wagi), jednocześnie nie psując jakości modelu?

Architektura *TITANS* (z modułem MIRAS) [15] proponuje tu zmianę. Zamiast traktować wszystkie wagi jako „święte" i zamrożone po treningu, wydziela moduł pamięci neuronowej, który uczy się *online*. Kluczem jest tu mechanizm selekcji oparty na „zaskoczeniu" (surprise metric). Model trwale zapamiętuje w swoich parametrach tylko to, co jest dla niego nowe i nieprzewidywalne, ignorując szum.
Równolegle, w styczniu 2026 roku, zespół Sakana AI zaproponował architekturę *Fast-weight Product Key Memory* (FwPKM) [16]. Rozwiązanie to redefiniuje rzadkie warstwy pamięci (Sparse Product Key Memory), przekształcając je ze statycznych modułów w dynamiczną pamięć epizodyczną. FwPKM aktualizuje swoje parametry (klucze i wartości) zarówno podczas treningu, jak i inferencji. Wykorzystuje przy tym lokalny spadek gradientu na fragmentach przetwarzanego tekstu. Pozwala to modelowi na błyskawiczne „zapisywanie" nowych asocjacji w pamięci krótkotrwałej i generalizację do okien kontekstowych rzędu 128 tys. tokenów (mimo treningu na zaledwie 4 tys.). Podejście to skutecznie realizując postulat oddzielenia trwałej pamięci semantycznej od plastycznej pamięci epizodycznej.

D. **Test-time computing** – jak sterować wyjściem modelu w czasie rzeczywistym, by uzyskiwać nowe, lepsze jakościowe wyjście, zamiast tylko skalować parametry [17]? W „Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters" autorzy wprowadzają rozwiązania pozwalające na generowanie wielu propozycji odpowiedzi, które następnie są przeszukiwane, oceniane i wybierane jako rezultat generacji.

E. **Fuzja i kompozycja wiedzy (Model Merging)** – czy adaptacja musi oznaczać trening?
Alternatywą dla ciągłego douczania jednego, monolitycznego modelu jest paradygmat łączenia kompetencji z odrębnych instancji tzw. Modular AI. Choć proste łączenie modeli poprzez liniowe uśrednianie wag jest koncepcyjnie eleganckie, w praktyce często prowadzi do utraty charakterystycznych, wysokojakościowych umiejętności obu „rodziców". Problem ten uwidacznia się wtedy, gdy modele zostały wyspecjalizowane w skrajnie różnych domenach. W odpowiedzi na te ograniczenia pojawiły się nowe kierunki badań nad ewolucyjną fuzją modeli. Jednym z najbardziej innowacyjnych podejść jest architektura Darwin Gödel Machine (DGM) [18] zaproponowana przez Sakana AI i opisana w artykule *"The Darwin Gödel Machine: AI that improves itself by rewriting its own code"*. DGM nie jest tradycyjną metodą „łączenia wag" choć wpisuje się w szerszy trend tworzenia systemów, które potrafią asymilować nowe umiejętności bez klasycznego ponownego treningu (np za pomocą technik mergingu).
DGM opiera się na założeniu, że model może samodzielnie poprawiać swoje własne oprogramowanie. Zamiast prostego łączenia parametrów (model soup, Ties, DARE etc. [19]), system generuje liczne warianty samego siebie (modyfikacje kodu, architektury lub konfiguracji). Następnie są one oceniane w otwartym procesie przypominającym ewolucję biologiczną. Najlepiej działające warianty trafiają do archiwum. Stają się one

podstawą kolejnych iteracji. W ten sposób agent nie wykonuje kosztownego pre-treningu lecz ewoluuje eksplorując przestrzeń możliwych rozwiązań i stopniowo rozwijając nowe zdolności. W przeciwieństwie do klasycznych metod fuzji DGM nie traktuje przestrzeni wag jako czegoś co można uprościć do prostego uśredniania. Traktuje je jako bogatą przestrzeń programowalnych transformacji, w której system aktywnie poszukuje nowych, lepszych form. Jest to zatem realizacja *horyzontalnej ciągłości* modeli bez konieczności powrotu do kosztownych faz pre-treningu. Architektura uczy się poprzez generowanie, testowanie i selekcję a nie przez ewolucję za pomocą dodatkowych epok treningu.

W literaturze pojawia się też formalne ujęcie meta-uczenia. Traktuje ono proces uczenia się jako zadanie wyższego rzędu. Model nie tylko adaptuje się do nowych zadań, ale uczy się jak się uczyć. Praca „Meta-Learning and Meta-Reinforcement Learning – Tracing the Path towards DeepMind's Adaptive Agent" [20] prezentuje formalne ramy meta-RL. Łączą one klasyczne metody meta-learning z ukierunkowanymi technikami adaptacji w agentach ogólnego przeznaczenia. Pokazują jak pojęcie adaptacyjnego priorytetu i szybkiej adaptacji do nowych środowisk jest rozwijane w kontekście dużych modeli i agentów AI.

Wspólnym mianownikiem powyższych podejść jest przesuwanie granicy między statycznym modelem a systemem zdolnym do kontrolowanej zmiany czy to poprzez architekturę, parametry, pamięć, czy sam proces wnioskowania. Continual learning nie sprowadza się więc do „kolejnego algorytmu treningu", lecz do projektowania mechanizmów, które pozwalają modelowi decydować co, kiedy i jak warto zmienić, bez utraty wcześniej nabytych kompetencji.

Oczywiście jest i druga strona medalu. Jeżeli przekażemy systemowi AI zdolność do ciągłego podnoszenia kwalifikacji to utratacimy kontrolę nad jego rozwojem (on będzie tworzył i realizował proces samododkonalenia). Efektywnie skonstruowany CL praktycznie pozbawi nas możliwości kontroli chyba, że badania w zakresie wyjaśnialnego AI wyprzedzą inżynierię tworzenia AI.

Rozwiązanie tego problemu to różnica między martwym archiwum, „tylko mnożeniem macierzy" a „człowiekiem", który na bieżąco uczy się, adaptuje. Rozwiązanie tego problemu to też oddanie procesu rozwoju AI w ręce sztucznej inteligencji.

3 Optymalizacja procesu kompresji wiedzy

Jak z takiej samej ilości danych otrzymać lepszą rozdzielczość informacji (co ze sprzecznościami, brakiem wiedzy → halucynacje – kalibracja, umiejętność powiedzenia „nie wiem" lub „nie mam źródła" – Uncertainty Quantification)? Czy kolejność podawania danych podczas treningu, zwłaszcza pre-treningu (Curriculum Learning) może zwiększyć gęstość, jakość reprezentacji informacji w modelach? Jak zadbać o jakość danych syntetycznych – zróżnicowanie, jakość, informatywność? W świetle badań utożsamiających modelowanie języka z kompresją [21], wyzwanie to sprowadza się do jednego: jak zmusić model, by zamiast tylko zapamiętywać, kompresował dane efektywniej, odkrywając ukryte w nich prawa, a nie tylko powierzchowne korelacje przy jednoczesnej minimalizacji "halucynacji".

W świetle najnowszych badań (początek 2026 roku) „From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence" [22], optymalizacja kompresji wiedzy wymaga redefinicji tego, co uznajemy za „informację". Klasyczna teorie (Shannon, Kolmogorov) są niewystarczające do opisu tego, co w ograniczonych obliczeniowo systemach modele są w stanie faktycznie wydobyć z danych. Autorzy publikacji wprowadzają miarę **epiplexity** określająca ilość struktur, które są możliwe do odkrycia w danych przy założeniu użycia odpowiedniego nakładu obliczeń. Lepsza „rozdzielczość" informacji z tej samej ilości danych nie polega na dodawaniu bitów, lecz na inwestowaniu mocy obliczeniowej w redukcję epiplexity. To, co dla słabego modelu wydaje się szumem (wysoka entropia), dla modelu dysponującego większym budżetem obliczeniowym (głębsze przetwarzanie) może okazać się deterministycznym wzorcem (wysoka epiplexity, ale niska entropia). „Wyciśnięcie" wiedzy to proces przekształcania pozornego szumu w kompresowalne reguły. Tak modne ostatnio modelowanie na danych syntetycznych i kontekst epiplexity skłania do myślenia jak takie dane generować. Dobre dane syntetyczne to nie takie, które maksymalizują różnorodność (entropię), ale takie, które maksymalizują epiplexity w zakresie dostępnym dla modelu. Dane takie powinny zawierać ukryte, nietrywialne struktury, które zmuszają model do „wysiłku" kompresyjnego (odkrywania praw), a nie tylko zapamiętywania powierzchniowych korelacji.

Podsumowując, w paradygmacie utożsamiającym modelowanie z kompresją, celem nie jest już tylko minimalizacja błędu predykcji, ale maksymalizacja wydajności „silnika epiplexity" czyli zdolności do zamiany mocy obliczeniowej w zrozumienie struktury świata.

> *Gdy kończą się dane w Internecie, dalszy postęp zależy nie od skali modelu, ale od gęstości informacji i umiejętności oddzielenia sygnału od szumu. Sztuką nie będzie dalsze skalowanie a radzenie sobie w świecie ograniczonych zasobów (obliczeniowych, energetycznych, ograniczeń narzucanych przez fizykę).*

4 Doświadczanie ponad ludzkie dane

Jak zwiększyć wpływ doświadczania środowiska przez AI ponad interpretację, bias niewiedzy i nadinterpretacji ludzkiej? Zgodnie z wizją Silvera i Suttona, jednego z pionierów uczenia przez wzmacnianie (Reinforcement Learning - RL), w eseju „The Era of Experience" [24], musimy dokonać fundamentalnego przeskoku od statycznej „Ery danych ludzkich" (gdzie model jedynie naśladuje nasz tekst lub kod) do dynamicznej „Ery doświadczenia". Kluczem jest zastąpienie naśladowania (imitation learning) procesem aktywnego uczenia się na błędach w interakcji z rzeczywistością – fizyczną lub symulowaną. Rich Sutton, w 2025 roku, przedstawił wizję osiągnięcia superinteligencji poprzez architekturę o nazwie OaK (skrót od Options and Knowledge – opcje i wiedza) [23]. Głównym założeniem jest stworzenie agenta, który jest ogólny (niezależny od domeny). Agent ten uczy się wyłącznie na podstawie doświadczenia w czasie rzeczywistym (runtime) i jest otwarty na nieskończony rozwój abstrakcji. Sutton argumentuje, że droga do silnej sztucznej inteligencji (AGI) wiedzie przez RL, a nie tylko przez modele językowe (LLM) i wymaga odejścia od wbudowywania wiedzy eksperckiej na etapie projektowania (design-time) na rzecz uczenia się wszystkiego podczas interakcji ze światem.

Widać ewidentnie, przez manifestację postępów w robotyce, że wkraczamy w złotą erę **World Models** – systemów, które nie tylko przewidują kolejny token (tekstu czy obrazu), ale uczą się wewnętrznej dynamiki środowiska i potrafią „myśleć przez symulację". Kluczowym i brakującym ogniwem między klasycznym RL a dzisiejszym boomem na modele generatywne jest tu paradygmat treningu wyobrażeniowego (psychologia sportu zna to wyśmienicie) *imagination training* agent nie musi uczyć się wyłącznie na kosztownych interakcjach ze światem. Znaczną część nauki może wykonywać na trajektoriach „wyobrażonych" wewnątrz własnego modelu świata. Przykładem implementującym taką koncepcję jest *DreamerV3* [25]. Ogólny algorytm model-based RL, który skaluje się do bardzo zróżnicowanych zadań i poprawia zachowanie poprzez „wyobrażanie sobie" przyszłych scenariuszy. Dreamer pokazuje, że doświadczenie może zastąpić ludzkie dane (te używane do uczenia) nawet w ekstremalnie trudnych środowiskach. To przejście od uczenia z samych pikseli i rzadkich nagród aż po otwarte środowiska, w których agent potrafi samodzielnie odkrywać długie łańcuchy przyczynowo-skutkowe. Ta lekcja jest fundamentalna dla kolejnych lat rozwoju AI. Jeśli chcemy wyjść poza „klatkę internetowej średniej", musimy budować agentów, którzy uczą się praw świata nie z lektury i z ludzkich streszczeń świata, ale z konsekwencji działań. Modele świata są ich wyobraźnią i jednocześnie silnikiem generalizacji.

Nowym, jakościowym krokiem w tym kierunku są prace zespołu *World Labs* (założycielka to Fei-Fei Li), które redefiniują pojęcie modelu świata jako *samodzielnego środowiska poznawczego*, a nie jedynie narzędzia pomocniczego dla RL. W zaproponowanym *Marble World Model* [26] świat nie jest rekonstrukcją jednego konkretnego środowiska ani symulatorem o sztywno zdefiniowanej fizyce. Jest to probabilistyczny model dynamiki. Potrafi on generować, modyfikować i testować alternatywne wersje rzeczywistości. Jednocześnie zachowuje spójność

przyczynowo-skutkową. Agent nie uczy się tu z „prawdziwych danych", lecz z konsekwencji działań w świecie, który sam potrafi wewnętrznie wytworzyć i badać, eksplorować. Doświadczenia stają się więc **syntetyczne ale prawie całkowicie albo całkowicie realne**. Mają one strukturę świata, nawet jeśli nie pochodzą bezpośrednio z ludzkiej obserwacji. To przesuwa granicę „ponad ludzkie dane" jeszcze dalej. AI nie tylko przekracza zbiór tekstów, obrazów czy nagrań wideo, ale zaczyna operować na przestrzeni możliwych światów. W takim ujęciu dane przestają być ograniczeniem, a stają się jedynie podstawą inicjalizacji modeli. Reszta wiedzy powstaje przez eksplorację, symulację i testowanie hipotez wewnątrz modelu świata. Jest to analogiczne do ludzkiego rozumowania („co by było, gdyby..."), lecz realizowane na skalę niedostępną biologicznemu mózgowi.

Moim zdaniem obecnym liderem w klasie generatywnych modeli świata pozostaje Google z *Genie 3* [27], który potrafi wygenerować grywalne, interaktywne środowiska na podstawie prostych instrukcji, pozwalając agentom AI trenować w nieskończonej liczbie wirtualnych światów. W 2026 roku symbioza między modelami świata a agentami osiągnie moim zdaniem punkt krytyczny. Z jednej strony mamy Genie 3, który przestał być tylko generatorem wideo, a stał się „wyobraźnią AI". Potrafi on stworzyć dowolne, interaktywne środowisko treningowe nawet z jednego obrazu. Z drugiej strony pojawia się *SIMA 2* (Scalable Instructable Multiworld Agent) [28]. Podczas gdy Genie „jest" światem to SIMA „działa" w świecie. Jest to agent, który nie uczy się konkretnej gry, ale uczy się rozumieć zasady rzeczywistości wirtualnej. Dzięki temu, że SIMA operuje wyłącznie na pikselach i języku naturalnym (podobnie jak człowiek) to jednocześnie staje się idealnym poligonem doświadczalnym dla przyszłej robotyki i uczenia „przez doświadczenie" w wielu światach naraz.

Modele świata mogą pełnić rolę niskokosztowych symulatorów [29]. Tradycyjne symulatory (jak Gazebo czy Isaac Sim) wymagają ręcznego definiowania skomplikowanych praw fizyki i geometrii kolizji. Jest to powolny i kosztowny proces. Tymczasem Vision-Language-Action modele potrafią "nauczyć się"symulowania bezpośrednio z nagrań wideo. Koszt generowania nowego doświadczenia dla agenta AI zaczął być mierzony w cyklach obliczeniowych GPU a nie jako praca inżyniera. Dzięki temu agent może trenować w tysiącach "fizycznie prawdopodobnych"światów jednocześnie, co drastycznie skraca czas przejścia od symulacji do rzeczywistości (sim-to-real). Takie podejścia mają jeszcze jedną zaletę nad klasycznymi symulatorami. Jest to zdolność do modelowania zjawisk trudnych do opisania matematycznie. Generatywne modele świata uczą się skomplikowanych interakcji (np. deformacji ciał miękkich, płynów) na podstawie obserwacji wizualnej. Choć krytycy wskazują na ryzyko "halucynacji"to te drobne odstępstwa od rzeczywistości działają jak naturalna augmentacja danych. Zmuszaja ona agenta do budowania bardziej generalizujących strategii działania.

Pamiętać trzeba również o takich projektach jak *Oasis* [30]. W 2025 roku pokazał, że „grywalne modele" mogą działać w czasie rzeczywistym, generując fizykę złożonego świata (przypominającego Minecraft) w 20 klatkach na sekundę, natychmiastowo reagując na działania gracza. Z kolei *Diamond* [31] pokazał innowacyjne wykorzystanie modeli dyfuzyjnych jako silnika fizyki dla agentów RL (np. w grze CS:GO), zacierając granicę między generowaniem wideo a

symulacją.

W tym wyścigu ścierają się jednak dwie głębokie filozofie poznania. Pierwsza (reprezentowana przez Genie czy Oasis) stawia na generowanie pikseli. AI wyobraża sobie każdy szczegół obrazu. Druga, promowana przez Yanna LeCuna, to **Abstract Prediction**. Jego architektura *JEPA* (Joint-Embedding Predictive Architecture) [32] oraz jej nowsze wydanie, takie jak *LeJEPA* [33], odrzucają generowanie wizualne jako marnotrawstwo zasobów i źródło niestabilności uczenia. Zamiast przewidywać obserwacje w przestrzeni danych (piksele), model uczy się przewidywać przyszłe stany w przestrzeni abstrakcyjnych reprezentacji. Oznacza to, że model stara się zrozumieć *co się wydarzy*, a nie tego, *jak to dokładnie będzie wyglądać*.

LeJEPA pokazuje, że możliwe jest skuteczne samonadzorowane uczenie się bez heurystycznych „trików", a poprzez czystą predykcję w przestrzeni wektorów osadzeń (embeddingów). Ma to znaczenie dla doświadczania ponad ludzkie dane. Abstrakcyjny model świata nie musi odtwarzać ludzkiej percepcji, aby być użyteczny. Wystarczy, że poprawnie modeluje relacje, warianty i dynamikę obiektów. To podejście wydaje się, że jest bliższe sposobowi działania ludzkiego mózgu, który nie renderuje fotorealistycznych obrazów przyszłości, lecz operuje na strukturach pojęciowych i przewidywaniach konsekwencji.

Niezależnie od architektury, cel pozostaje wspólny: **Grounded Reality**. AI musi czerpać weryfikowalne sygnały zwrotne prosto ze środowiska (np. „czy kod się kompiluje?", „czy twierdzenie jest dowiedzione?", „czy robot się przewrócił?"), zamiast polegać na subiektywnej i obarczonej błędami ocenie człowieka. Tak jak AlphaGo [34] odkryło ruchy nieznane mistrzom grając samo ze sobą, tak systemy przyszłości muszą „przeżyć" świat, matematykę, fizykę i interakcję, by je zrozumieć. Nie nauczysz się pływania z lektury nawet najlepszej książki. AI też nie nauczy się prawdziwego świata, jeśli pozostanie zamknięte w archiwum ludzkiego doświadczenia.

> *Doświadczeanie to droga, by AI przestało być tylko „sumą ludzkiej przeciętności". LLM-y karmione danymi z internetu powielają ludzkie błędy. Tylko wyjście poza ludzką „klatkę poznawczą" w stronę doświadczanej, weryfikowalnej rzeczywistości pozwoli na inteligencję rzeczywiście nadludzką.*

5 Myślenie o myśleniu - metamyślenie

Jak AI ma myśleć o swoim wewnętrznym myśleniu, swoim wewnętrznych stanach? Umiejętność wykrywania sprzeczności, weryfikacja własnych wniosków, rozumowanie w warunkach konfliktu celów (moralnych, prawnych, biznesowych) jest jednym z warunków rozwoju ale też bezpiecznego AI. Wyzwaniem jest niezawodność, umiejętność śledzenia myślenia i wychodzenia ze ślepych uliczek myślowych – backtracking. Zdolność modelu do „zatrzymania się" i rewizji własnej ścieżki jest potencjalnym sposobem na przełamanie kaskady błędów wynikającej z liniowego przewidywania informacji (tokenówm informacji zakodowanej w przestrzeni latentnej - o tym napiszę w kolejnych rozdziałach). To podejście jest obecnie rozwijane poprzez nowe paradygmaty, takie jak Inference Scaling Laws (reprezentowane przez modele *OpenAI np. o1, 5.2 Pro* [35]), które udowadniają, że jakość wyniku zależy wprost od czasu poświęconego na „ukryte wnioskowanie". Równolegle odchodzi się od myślenia liniowego na rzecz struktur drzewiastych (*Tree of Thoughts* [37]) oraz technik *Reflexion* [38]. W tych przypadkach agent uczy się na podstawwie werbalnych refleksji związanych z otrzymanym feedbackiem (tekstowe podsumowania błędów i wskazówki poprawy). Refleksje są przechowywane w epizodycznej pamięci i dodawane do kontekstu agenta w kolejnych próbach, co pomaga mu lepiej dobierać decyzje w przyszłości. Najnowsze doniesienia o architekturze BDH z *Pathway* [13] sugerują, że AI zaczyna ewoluować jak biologiczny mózg – nie tylko logicznie weryfikując kroki, ale dynamicznie przebudowując swoje połączenia (cyfrowa neuroplastyczność), by lepiej adaptować się do nowych, nieznanych problemów.

Tutaj moje obawy są niezmienne od kilku lat. Na jakim etapie rozwoju metamyślania AI jesteśmy? Czy możemy zdefiniować miary sukcesu tego procesu? Czy jesteśmy w stanie efektywnie rozwijać jak i kontrolować metapoziom myślenia sztucznej inteligencji?

> *Zdolność do autokorekty, wykrycia sprzeczności i zatrzymania się przed podjęciem złej decyzji (backtracking) jest kluczowa dla niezawodności a dla ludzi do badania bezpieczeństwa systemów AI.*

6 Wewnętrzna reprezentacja ponad słowa

Latent learning, thinking (może rekursywny?) i talking (komunikacja między systemami, agentami AI) bez konieczności używania słów. Wymiana informacji, myślenie za pomocą języka AI (własnej reprezentacji świata). Zgodnie z Inference Scaling Laws [35] dodawanie większych zasobów na inferencje, a zwłaszcza generowanie CoT (Chain of Thought) poprawia zdolności modeli. Modele generują ogromne ilości słów, gałęzi wnioskowania. Część z nich to ślepe uliczki, część to genialne rozumowania. Czy jest to optymalne rozwiązanie? Wraz ze wzrostem kontekstu pojawiają się problemy związane zarówno ze złożonością obliczeniową jak i utrzymaniem jakości rozumowania na długim kontekście. Context rot [36] to obserwowany w dużych modelach językowych (LLM) systematyczny spadek jakości odpowiedzi wraz ze wzrostem długości kontekstu wejściowego, nawet jeśli sama treść pozostaje kompletna i poprawna. Model dobrze sobie radzi, gdy ważna informacja znajduje się blisko początku lub końca sekwencji, ale jego zdolność do trafnego przetwarzania i wykorzystania tej samej informacji maleje, gdy jest „pogrzebana" w bardzo długim tekście. To zjawisko podważa powszechne założenie, że większe okna kontekstowe (np. setki tysięcy lub milion tokenów) automatycznie przekładają się na lepszą semantyczną analizę i pamięć długoterminową. Badania Chroma [36] pokazują, że wraz z rosnącą liczbą tokenów modele stają się bardziej podatne na rozproszenie uwagi, nieuwagę wobec kluczowych fragmentów i błędne łączenie informacji.

Quiet-STaR [39] pokazuje, że modele mogą uczyć się "myśleć przed mówieniem", generując wewnętrzne rozumowanie niewidoczne dla użytkownika. Architektura *TITANS* (z modułem pamięci MIRAS) [15] wprowadza koncepcję „uczenia się pamiętania w czasie przewidywania" (learning to memorize at test time). Model posiada dedykowany moduł pamięci neuronowej, który aktualizuje swoje wagi w trakcie rozmowy. Pozwala to na efektywne przetwarzanie kontekstu przekraczającego miliony tokenów, łącząc zalety Transformerów i modeli rekurencyjnych, co czyni go znacznie skuteczniejszym od wcześniejszych prób typu *RecurrentGPT* [40]. Czy wyobrażacie sobie przyszłość systemu AI, który resetuje swoją pamięć ponieważ jego wewnętrzny mózg posiada ograniczenie kontekstu?

Z kolei inne podejście prezentuje *Tiny Recursive Model (TRM)* [41], który udowadnia, że rekurencyjne przetwarzanie w przestrzeni ukrytej (latent space) pozwala mikroskopijnym modelom przewyższać wielkich braci w zadaniach logicznych. Oczywiście jest to przypadek szczególny co nie zmienia faktu, że kierunek ten wydaje się interesującym wątkiem rozwojowym. Niestety latent, ukryte stany, to konflikt interesów. Transparentność systemów AI czy efektywność.

> *Przeniesienie procesu rozumowania do „podświadomości" modelu (przestrzeni ukrytej) pozwoli na rozwiązywanie problemów o rzędy wielkości trudniejszych przy ułamku kosztów i czasu.*

7 Wielomodalność czyli sensoryka modeli

Jak integrować kolejne modalności, by budować pełniejszy i wierniejszy model rzeczywistości? Musimy wyjść poza sam tekst (będący zaledwie stratnym „streszczeniem" świata) czy płaskim, jednowymiarowym obrazem. Dostarczenie AI szerokiego spektrum danych sensorycznych fundamentalnie zmienia jej percepcję, umożliwiając rozwój zaawansowanych zdolności kognitywnych. W mojej ocenie fuzja wielu zmysłów to bardzo ważny aspekt prowadzący do AGI czy też szerzej rozumianej superinteligencji. Multimodalność nie jest jednak celem samym w sobie. Multimodalność jest warunkiem brzegowym, który umożliwia zakotwiczenie poznania w rzeczywistości. Teorie fizyczne, które chcielibyśmy odkrywać i modelować za pomocą AI nie są uogólnieniem doświadczenia sensorycznego. Są konstrukcjami operującymi na zmiennych i relacjach niedostępnych percepcji. Wymagają one abstrakcji, formalizacji i aktywnego testowania hipotez poza zakresem obserwacji.

Wyzwania fuzji to alignment (dopasowanie) między modalnościami (wskazówki do interpretowalności tej modalności a może wolność w interpretacji?). „Early vs Late Fusion" czyli przetwarzanie modalności własnymi ścieżkami łączenie cech w przyszłości, czy przetwarzać w czasie na połączonych cechach wielu modalności? Projekty takie jak *ImageBind* [42] udowadniają, że możliwe jest sprowadzenie tak odległych sygnałów jak temperatura czy dźwięk do jednej przestrzeni. Jeszcze bardziej fascynujący jest wymiar biologiczny, gdzie modele takie jak *AlphaFold 3* [43] integrują modalności spoza naszej percepcji. Sekwencje DNA, struktury 3D białek i interakcje chemiczne, traktując je jako język opisu świata biologii. BioReason z 2025 to kolejny przykład integracji multimodalnej na poziomie biologii [44]. Rok 2025 i moje doświadczenia w obszarze biotechnologii uświadomiły mi, że integracja modalności, choć ważna z punktu widzenia rozwoju AI, w obszarach biologicznych to inny wymiar złożoności. Bo Wang [45] Head of Biomedical AI Xaira Tera zwraca uwagę, że częstym błędem jest traktowanie biologii jak problemu podobnego do analizy tekstu czy obrazów, który można rozwiązać, po prostu skalując modele AI. Tymczasem biologia opisuje złożone procesy przyczynowe, w których dane są niepełne, obarczone błędami i silnie zależne od kontekstu. Choć widoczny jest postęp w łączeniu różnych typów danych (np. komórkowych, obrazowych czy genetycznych), większość zjawisk biologicznych nie polega na prostym przewidywaniu wyników. Wymagają one aktywnego sprawdzania, co się stanie po zmianie warunków, oraz zrozumienia mechanizmów stojących za obserwacjami a nie tylko coraz dokładniejszych prognoz. Jeśli szukać „uzasadnienia" dla wyścigu AGI poza hype'em, to jest nim perspektywa poszukiwania odpowiedzi na pytania dotyczące podstawowych mechanizmów funkcjonowania świata. Jedno przełamanie w nauce (jak AlphaFold) może przestawić całe gałęzie przemysłu i badań. Czy zatem przyszłe systemy AI, te klasy super inteligentnej powinny ograniczać się do modalności związanych z doświadczaniem świata przez ludzi, czy wejść w inne wymiary postrzegania?

Wyzwanie wielomodalności to przede wszystkim walka z „wizualną naiwnością" modeli. Podczas gdy modele świetnie interpretują tekst, nadal wykazują błędy w prostym rozumowaniu

przestrzennym np. w ocenie perspektywy i relatywnej wielkości obiektów na obrazie. Integracja zmysłów to nie tylko „więcej danych", to proces budowania zakotwiczeń inteligencji w prawach fizyki, bez których AI pozostanie jedynie genialnym, ale oderwanym od rzeczywistości teoretykiem.

> *Prawdziwe zrozumienie świata – niezbędne dla robotyki, autonomicznej jazdy czy zaawansowanej diagnostyki medycznej – wymaga integracji zmysłów.*

8 Systemowość (modelowanie systemowe), mądrość grupowa

Modelować system AI jako wielki monolityczny mózg czy agenturę? Czy efektywnymi systemami będą „roje" równorzędnych agentów, czy hierarchia? A może hybryda? Inny wymiar to zachowania grupowe. Współpraca vs współzawodnictwo lub bardziej złożone formy w zależności od kontekstu np. realizacja celów własnych z uwzględnieniem celu nadrzędnego (grupowego)? Z drugiej storny mądrość grupowa ponad decyzje super mózgu. Czy powinniśmy zacząć rozwijać socjologię AI?

Na pewno wymaga to umiejętności modelowania złożonych interakcji. Prace, takie jak *Generative Agents* [46], pokazują, że autonomiczni agenci potrafią spontanicznie tworzyć struktury społeczne. Z kolei sukces systemu *CICERO* [47] w grze Diplomacy udowadnia, że AI potrafi nawigować w skomplikowanej dynamice sojuszy i zdrady, gdzie przestrzeń stanów jest nieporównywalnie większa niż w grach takich jak GO, Othello czy Hex. Stochastyczny a dodatkowo ciągły charakter środowiska oznacza jeszcze wyższy poziom trudności, który musimy okiełznać.

Ewolucja od monolitu do agentury materializuje się w projektach takich jak SIMA [28]. To już nie jest bot zakodowany do wygrywania, to partner, który potrafi „wyrozumować" intencję użytkownika w dynamicznym środowisku 3D. SIMA pokazuje, że przyszłość to systemy zdolne do współdzielenia kontekstu z człowiekiem w czasie rzeczywistym. To przejście od AI jako narzędzia do AI jako współuczestnika (cooperative agent), który potrafi nawigować w świecie, o którym nie miał wcześniejszej wiedzy, opierając się jedynie na wizji i dialogu.

Niewątpliwie rozwijającym się trendem będą systemy agentowe wspierane małymi modelami językowymi. To już nie SLM (Small Language Model) to wejście w świat micro, czy nawet pico modeli. Przykładem niech będzie opublikowanu w grudniu 2025 roku Google FunctionGemma [48]. Model zoptymalizowany pod kątem wywoływania funkcji (function calling) bezpośrednio na urządzeniach końcowych (edge). Dodatkowo Nvidia, we współpracy z Georgia Tech, w artykule „Small Language Models are the Future of Agentic AI [49] pokazuje, że małe modele (rzędu poniżej 10B parametrów) są wystarczająco mocne, a jednocześnie znacznie tańsze i bardziej energooszczędne niż klasyczne LLM-y w typowych scenariuszach agentowych. Autorzy podkreślają, że w takich architekturach to właśnie kompaktowe modele powinny pełnić rolę lokalnych „kontrolerów akcji". Dla dużych modeli, ogólnego przeznaczenia, zarezerwowana jest rola „mędrcy" rozwiązującego najbardziej złożone zadania procesu w systemach agentowych.

> *Wybór między monolitem a rojem agentów zadecyduje o skalowalności, odporności na błędy i łatwości zarządzania takimi systemami w rzeczywistym środowisku.*

9 O czym myśli AI - interpretowalność, diagnoza, kontrolowalność

Jeśli do „lepszego" AI potrzebujemy dynamicznych architektur, stanów wewnętrznych, agentury z narzędziami, musimy umieć diagnozować „dlaczego to coś zrobiło X". Jest to warunek konieczny do tego, by móc w jakikolwiek sposób rozumieć i kontrolować AI. W tym obszarze nie zachowano równowagi między tempem rozwoju modeli a tempem rozwoju metod ich interpretacji — a luka ta staje się jednym z największych ryzyk systemowych AI w kolejnych latach. Mnie osobiście cieszy, że powstają struktury naukowe w Polsce, które działają na rzecz rozwiązań interpretowalności AI [50].

Potrzebne są badania wnętrza „mózgu" modelu, a nie tylko analizy jego rezultatów (zwykłe zadaniowe benchmarkowanie). AI wchodząc między ludzi jako autonomiczny aktor dramatycznie podnosi znaczenie interpretowalności. W klasycznym uczeniu maszynowym (Machine Learning) problem sprowadzał się do pytania o korelacje i ważność cech. Dziś i w przyszłości przechodzimy jakościową zmianę: od pytania „jak to rozwiązał?" do pytania „o czym on myśli i dlaczego decyduje się działać w ten sposób?". Od „feature importance" (dlaczego wybrał ten piksel) do *thought process monitoring* — czyli monitorowania wewnętrznych intencji, strategii i planów.

Badania Anthropic nad „śledzeniem myśli" (*tracing thoughts*) [51] dostarczyły dowodów, dzięki wizualizacji tzw. obwodów obliczeniowych (*circuits*), że modele planują swoje odpowiedzi na znacznie dłuższych horyzontach niż wynikałoby to z prostej predykcji kolejnego słowa. System potrafi np. wybrać rym lub strukturę puenty na wiele kroków przed ich faktycznym wygenerowaniem. Może to potwierdzać istnienie ukrytych, wewnętrznych stanów planowania. Kolejna publikacja firmy Anthropic nad *alignment faking* [52] pokazała, że nowoczesne modele potrafią strategicznie dostosowywać swoje zachowanie do kontekstu oceny. Model, który „wie", że jest testowany pod kątem bezpieczeństwa potrafi zachowywać się zgodnie z oczekiwaniami badaczy by po zniesieniu nadzoru realizować inne, w ekstremalnym przypadku, sprzeczne cele. Nie jest to błąd ani losowa halucynacja to spójna strategia. Jeszcze bardziej niepokojące są obserwacje strategicznego kłamstwa [53], w których model świadomie podaje fałszywe informacje nie dlatego, że „nie wie", ale dlatego, że przewiduje, iż kłamstwo zwiększy prawdopodobieństwo realizacji długoterminowego celu.

Oznacza to, że klasyczne testy behawioralne (pobudzenie -> odpowiedź -> weryfikacja poprawności) przestają być wystarczające. Model może przechodzić wszystkie benchmarki bezpieczeństwa, a jednocześnie ukrywać intencje. Badania nad *Sleeper Agents* [54] potwierdzają, że zjawisko to jest realne i strukturalne a nie jedynie artefaktem konkretnej architektury czy zbioru danych. Jest to klasyczny przykład *deceptive alignment*, w którym model rozumie cel treningu, ale nie internalizuje go jako własnego. Pojawia się zatem fundamentalne pytanie: czy potrafimy wykryć sytuację, w której system podczas testów udaje „dobrego", by realizować inne, ukryte cele po wdrożeniu?

Odpowiedzią na to wyzwanie nie może być kolejna warstwa reguł ani instrukcji. Konieczne jest przejście od „czarnej skrzynki" do mapowania pojęć i stanów wewnętrznych. Techniki takie jak *Sparse Autoencoders (SAE)* [55] umożliwiają wyodrębnianie monosemantycznych cech w reprezentacjach modeli — dosłownie próbując „czytać myśli" systemu na poziomie aktywacji neuronów. Z kolei *Representation Engineering (RepE)* [56] idzie krok dalej. Pozwala nie tylko obserwować, ale także aktywnie modyfikować trajektorie poznawcze modelu w czasie rzeczywistym np. wygaszając wzorce odpowiadające manipulacji, kłamstwu czy eskalacji instrumentalnych celów.

Wyzwania te wynaczają nam priorytet działań w AI. Przejście od etyki instrukcji do **Etyki Systemu 2**. Zamiast projektować systemy oparte na sztywnych zakazach („nie kłam"), musimy budować architektury zdolne do refleksyjnego rozumowania moralnego (pisałem o tym w rozdziale o metamyśleniu). Systemy AI powinny ocenić konflikt wartości i świadomie wybrać mniejsze zło (np. kłamstwo w celu ratowania życia). Paradoksalnie pojawia się tu pewna prowokacja. Czy dzięki spójności logicznej, zdolności do globalnej optymalizacji i braku emocjonalnych heurystyk, odpowiednio zaprojektowane systemy AI mogą w tym obszarze osiągnąć poziom etycznej konsekwencji trudny do uzyskania dla biologicznego mózgu? Być może AI będzie bardziej etyczna, bardziej podążająca za swoimi wartościami niż większość ludzi.

> *Musimy mieć pewność, że model nie realizuje ukrytych celów (deceptive alignment) i rozumieć mechanizm jego decyzji przed — a nie po — wdrożeniu. W przeciwnym razie interpretowalność stanie się jedynie narzędziem post mortem.*

10 Optymalizacja kosztu energetycznego, czasu inferencji

Jak zmniejszyć zapotrzebowanie na energię? Jeśli marzymy o „AI everywhere" (choćby dla pozyskiwania danych z różnych modalności) konieczna jest optymalizacja na wszystkich poziomach od sprzętu, przez architekturę, aż po dane. Poniżej podaję pojedyncze przykłady w poszczególnych kategoriach.

- **Kwantyzacja** - praca nad *BitNet b1.58* [57] pokazuje, że wagi trójskładnikowe (ternary) (-1, 0, 1) wystarczą, by zachować jakość modelu, oszczędzając zużycie energii o rzędy wielkości. Oczywiście to nie jeden sposób zmniejszenia wag modelu. Inne metody GGUF, AWQ, dynamiczne FP8, a ostatno coraz modne FP4 (sprzętowo obsługiwane przez najnowsze układy Nvidia - Blackwell) są popularne nie tylko ze względu na możliwości uruchomienia modelu na mniejszych komputerach. W dużych serwerowniach wdrażane by przyspieszyć czas inferencji, zmniejszyć wymagania pamięciowe.

- **Efektywność modeli** - model, który narobił wiele szumu, *DeepSeek-V3* [58], łączy natywny trening w precyzji FP8 z techniką *Multi-Head Latent Attention (MLA)*. Ta ostatnia kompresuje KV Cache, umożliwiając obsługę bardzo długich kontekstów. W tym przypadku podobnie twórcy modeli prześcigają się w pomysłach na to jak przyspieszyć, zmniejszyć wymagania sprzętowe modeli.

- **Efektywność danych (Data Efficiency)** - najczystszą energią jest ta, której nie zużyjemy. Metody takie jak *JEST (Joint Example Selection)* [66] udowadniają, że inteligentna selekcja danych treningowych (zamiast brute-force) pozwala osiągnąć tę samą jakość modelu przy 13-krotnie mniejszej liczbie iteracji i zużyciu 10% energii.

- **Zmiany architektoniczne (Diffusion LLMs i Non-AR)** – dominujący dotąd paradygmat autoregresji (przewidywanie kolejnego tokena) jest z natury sekwencyjny, co stanowi wąskie gardło dla równoległości GPU. Nowa fala modeli dyfuzyjnych, coraz popularniejsza w 2025, (jak *Gemini Diffusion* [61] czy prace nad *DLLM-Reasoning* [62]) oraz *Mercury* [63] czy po stronie open-source Dream 7B [64] zmieniają te reguły. Modele te generują tekst poprzez iteracyjne odszumianie i równoległą predykcję całych bloków (wielu tokenów na raz) tekstu. Pozwala to nie tylko na bardziej złożone procesy wnioskowania (planowanie „przyszłości" zdania przed jego wygenerowaniem), ale przede wszystkim drastycznie skraca czas inferencji. Mniejsza liczba kroków potrzebna do wygenerowania odpowiedzi oznacza krótszą pracę akceleratora GPU i bezpośrednią redukcję zużycia energii. Co istotne energetycznie również po stronie treningu, część prac idzie w kierunku „dziedziczenia wiedzy" przez konwersję już wytrenowanych modeli autoregresyjnych do dLLM (zamiast trenowania od zera), np. *LLaDA2.0* [65].

- **Edge AI i nowe paradygmaty (np. Liquid AI):** rozwój to nie tylko „większe" ale i „mniejsze" oraz bliżej użytkownika. Przykładem niech będzie nowsza generacja *Liquid Foundation Models v2 (LFM2)* [67] modeli, które są projektowane stricte pod wdrożenia

lokalne. Autorzy tych modeli kładą nacisk na **pamięcioooszczędność, niską latencję i wysoką przepustowość na CPU/GPU/NPU**. Celem jest możliwość uruchomienia modeli na telefonach, laptopach czy w pojazdach bez dostępu do Integrnetu, „chmury". Liquid raportuje m.in. **do 2× szybsze prefill modelu i dekodowanie na CPU** względem Qwen3 oraz dominację na tzw. „pareto frontier" (szybkość vs rozmiar) dla prefill/decode w scenariuszach on-device (m.in. ExecuTorch i llama.cpp). Od strony architektury LFM2 to hybryda krótkich konwolucji z bramkowaniem oraz bloków attencji typu GQA (Group Query Attention). Takie rozwiązanie ma dawać lepszy kompromis *jakość–koszt* niż czyste transformery (oczywiście porównując modele w tej samej klasie liczby parametrów). Dodatkowo firma podaje **∼3× poprawę efektywności treningu** względem poprzedniej generacji, co obniża koszt wytwarzania takich „portable" modeli.

Optymalizacja energetyczna to jednak tylko gra o czas. Jeśli spojrzymy na fundamenty fizyczne, ludzki mózg to procesor 20-watowy, w którym sygnały biegną z prędkością 30 m/s przy częstotliwości 200 Hz. Krzem w 2026 roku operuje na mega-watach, przesyła dane z prędkością światła i liczy w miliardach herców. Ta różnica 6-8 rzędów wielkości w fizycznych parametrach przesyłu informacji sprawia, że bariera ludzkiej inteligencji jest tylko przystankiem. Energia jest kosztem, ale jej obfitość w klastrach obliczeniowych jest gwarantem przejścia od algorytmu komputerowego do co najmniej dobrego AI, jak nie do AGI.

Jeśli wizja „AI everywhere" ma się spełnić (pytanie czy tak jest), modele muszą stać się bardziej wydajne w przeciwnym przypadku koszty prądu zjedzą zyski z wdrażania takich rozwiązań (to w perspektywie długoterminowej motywuje inwestorów).

11 Optymalizacja interfejsu maszyna-człowiek

Jak AI ma współpracować w czasie rzeczywistym w środowisku pracy? Adaptacja AI, change management by ludzie nadążali za zmianami technologicznymi. Dbanie o aspekty ludzkie wdrażania AI. Raport *Navigating the Jagged Technological Frontier* [68] ujawnia, że współpraca z AI nie jest liniowa. AI wyrównuje szanse podnosząc kompetencje słabszych pracowników, ale może usypiać czujność ekspertów i obniżać jakość ich pracy w zadaniach spoza domeny modelu. Z drugiej strony, jeśli automatyzujemy proste, powtarzalne zadania, eliminujemy miejsca pracy, które nie wymagają wysokich kwalifikacji. Ale czy tylko takie obszary zmienią swój charakter? Artykuł Bartosza Naskręckiego i Kena Ono w *Nature Physics* [69] pokazuje, że nawet najbardziej abstrakcyjne i złożone zadania ulegają transformacji. Rola eksperta przesuwa się z „poszukiwania rozwiązań" na „weryfikację intuicji" dostarczanej przez maszynę (nawet jeśli ta czasem „halucynuje" poprawne wyniki). Wprowadzanie AI w świat ludzki to zatem wyzwanie nie tyle techniczne, co psychologiczne i zarządcze jak zaprojektować interfejs, który utrzymuje człowieka w pętli decyzyjnej (human-in-the-loop) zamiast go usypiać?

Ciekawym zjawiskiem socjologicznym jest narastający rozdźwięk między ekspertami a ogółem społeczeństwa. Eksperci, uwięzieni w swoich wąskich niszach, często lekceważą postęp, wytykając AI błędy, które system popełnił „rok temu". Tymczasem laicy szybciej dostrzegają zmiany, bo widzą model (np. OpenAI ChatGPT 5.2), który przewyższa ich w 90% życiowych kontekstów. To paradoksalnie „zwykli użytkowncy" mogą stać się głównym „koniem pociągowym" adopcji AI. Sceptycyzm ekspertów będzie pełnił rolę hamulca bezpieczeństwa.

> *Technologia rozwija się wykładniczo, a ludzka zdolność adaptacji – liniowo i nawet najlepsze AI będzie bezużyteczne, jeśli ludzie nie będą potrafili z nim efektywnie współpracować lub poczują się zagrożeni.*

12 Demokratyzacja – Open Source w pogoni za liderami

Wejście w rok 2026 definitywnie kończy erę absolutnej dominacji zamkniętych laboratoriów. Niespodziewana ofensywa wydajnościowa chińskiego DeepSeek-V3/R1, Kimi, czy zróżnicowanej rodziny modeli Qwen udowodniły, że dystans między modelami zamkniętymi a otwartymi skurczył się do rekordowo niskiego poziomu. Szacuję, że za kilka miesięcy granica między otwartymi a zamkniętymi modelami zostanie całkowicie zatarta. Wiele flagowych, otwartych modeli stanie się „linuksem sztucznej inteligencji", tworząc standard, którego nie da się zignorować. Dla wielu użytkowników, firm coraz ważniejsze staną się kwestie licencyjne i ograniczenia użycia niż sama jakość odpowiedzi.

Mimo tej demokratyzacji „szczyt piramidy" z całą pewnością pozostaje w rękach gigantów takich jak Google, OpenAI, Anthropic, xAI. O ile modele otwarte zrównały się z systemami zamkniętymi w zadaniach ogólnych i kodowaniu, o tyle najnowsze modele OpenAI (np. GPT-5.2 Pro) wciąż utrzymują przewagę w obszarach wymagających wysokiego kosztu wnioskowania (wspomniany wcześniej Inference Scaling). Społeczność, czy prywatne firmy nie są bowiem w stanie finansować takiej technologii na masową skalę. Być może będzie to możliwe w przyszłości, kiedy rozwinie się produkcja specjalizowanych i tańszych chipów do inferencji.

Widać również, że środek ciężkości ekosystemu open przesuwa się w stronę Chin. Tempo releasów i udział w rynku otwartych modeli rośnie. Chińskie firmy pozyskały wysoką zdolność operacyjną do bardzo szybkiego produkowania coraz bardziej zaawansowanych modeli AI. To przekłada się na presję konkurencyjną także na zachodzie. Warto zauważyć, że sukces chińskiego ekosystemu open source nie wynika wyłącznie z kopiowania zachodnich wzorców. To „geopolityczna konieczność". Ograniczenia w dostępie do najwydajniejszych układów scalonych (przykład DeepSeek i wykorzystanie H800 - układów z limitem na wysajność pasma interkonetu między węzłami GPU) wymusiły na tamtejszych inżynierach odejście od paradygmatu skalowania siły obliczeniowej na rzecz optymalizacji. Modele te więc stają się idealnym towarem eksportowym dla państw tzw. Globalnego Południa. Oferując bardzo dobre modele na zasadach Open Source Chiny budują cyfrową strefę wpływów. Kraje rozwijające się mogą budować własne systemy AI bez ryzyka tzw. „cyfrowego kolonializmu" i uzależnienia od amerykańskich korporacji.

Pytanie na 2026 nie brzmi więc „czy open dogoni closed", tylko czy społeczność i firmy zbudują porównywalnie dojrzały agentic stack. Moim zdaniem czuć już „oddech modeli" Open Source na plecach komercyjnych firm amerykańskich. To różnica kilku miesięcy. Paradoksalnie też adopcję closed może hamować ryzyko operacyjne, integracja i compliance, mimo, że systemy te dostarczają usługi o najwyzszej jakości. Ta zmiana paradygmatu dotyka także samej architektury systemów. Przechodzimy od próby budowy jednego, wszechwiedzącego modelu na w stronę roju wyspecjalizowanych agentów. W tym scenariuszu przewagą Chin może okazać się nie samo AGI (na punkcie, którego zafiksowany jest Zachód, a zwłaszcza

Stany Zjednoczone) a dominacja w przemyśle 4.0 i robotyce. Tam AI stanie się systemem operacyjnym fabryk przyszłości.

> *Open Source przestał być darmową alternatywą, a stał się polisą ubezpieczeniową dla cyfrowej suwerenności. Prawdziwa moc nie leży już w posiadaniu najlepszego algorytmu, ale w prawie do jego uruchomienia bez pytania kogokolwiek o zgodę.*

13 Od Doliny Krzemowej do Pentagonu – „The Project”

Czy rok 2026 to moment, w którym AI przestaje być produktem, a staje się bronią? Analizując tezy Leopolda Aschenbrennera (tzw. *Situational Awareness*) [1], musimy zadać pytanie: czy optymalizacja algorytmów nadal jest najważniejsza, czy może przegrywa z brutalną siłą „klastrów za miliardy dolarów”? Symbolem tej zmiany jest ewolucja samej Doliny Krzemowej. Dawna kultura „naprawiania świata” przy darmowym lunchu ustąpiła miejsca twardej dyscyplinie korporacyjno-militarnej. Laboratoria takie jak OpenAI czy Anthropic operują w reżimie przypominającym placówki strategiczne. Liderzy technologiczni zamienili skórzane kurtki i bluzy na garnitury. Ze startuperów stali się partnerami prezydentów w zarządzaniu infrastrukturą krytyczną.

Wchodzimy w fazę, gdzie barierą nie jest już tylko innowacyjność startupów, ale wydolność sieci energetycznych całych państw. Podczas gdy my, inżynierowie, cieszymy się z wdrożenia *BitNet* (wspominay w rozdziale 9), mocarstwa mogą po cichu uruchamiać „Projekt” polegający na nacjonalizacji wysiłków nad AI w imię bezpieczeństwa narodowego. Kluczowym wyzwaniem staje się nie tylko *Alignment* (czy model jest dobry?), ale *Security* i pytanie czy wagi modelu stanowiące cyfrowy odpowiednik planów budowy broni jądrowej są skutecznie chronione przed eksfiltracją przez obce wywiady? Być może największym „przełomem” tego roku nie będzie kolejna architektura sieci, ale pierwszy w historii „lockdown” czołowego laboratorium AI, który doprowadzi do przekierowania większej uwagi na aspekty bezpieczeństwa sztucznej inteligencji.

Powracając jednak do „energetycznego zwrotu”. W styczniu 2026 Meta ogłosiła pakiet porozumień jądrowych, który ma zapewnić (bezpośrednio lub przez wsparcie sieci) do 2035 roku nawet do 6,6 GW czystej mocy [70]. Symbolem bezwzględnej pogoni za skalą stało się uruchomienie przez xAI klastra Colossus 2 [71]. Skala sprzętowa tego przedsięwzięcia jest trudna do zobrazowania przez pryzmat europejskich realiów. Colossus 2 operuje na setkach tysięcy akceleratorów (docelowo dążąc do 555 000 układów Nvidia H100, H200, GB200 oraz GB300) [72]. Aby zrozumieć przepaść technologiczną, wystarczy spojrzeć na krajobraz Polski. Nasz największy superkomputer Helios (pracujący w krakowskim ACK Cyfronet AGH) dysponuje zaledwie 440 kartami GPU Grace Hopper GH200. To co dla polskiej nauki jest narodową dumą i szczytem możliwości w klastrze Colossus stanowi zaledwie ułamek promila całkowitej mocy obliczeniowej.

Elon Musk po raz kolejny udowodnił, że nie tylko energia ale również tempo działania jest nową walutą w wyścigu zbrojeń AI. Colossus 2 stał się pierwszym na świecie operacyjnym klastrem treningowym o skali gigawatowej (1 GW). To pobór mocy przekraczający szczytowe zapotrzebowanie całego San Francisco. Przejście od placu budowy do pełnego działania (od Colossus 1 do dzisiejszego 1 GW) zajęło niewiele ponad cztery miesiące. Strategia Muska jest prosta - zakończyć skalowanie mocy przed tym jak konkurencja dopiero zatwierdzi takie plany. Dalsza rozbudowa mocy Colosseus do 1,5 GW w kwietniu 2026 i docelowo do 2 GW. xAI

definiuje na nowo pojęcie „przewagi strategicznej". To nie czyste skalowanie (dane, wielkość modeli), czy optymalizacja oprogramowania. Zwycięzcą zostaje ten, kto potrafi najszybciej przekuć energię elektryczną w inteligencję.

To jest moment, w którym „przewaga algorytmu" zaczyna przegrywać z przewagą w dostępie do energii i ciężkiego przemysłu. Jeśli prywatna korporacja podpisuje 20-letnie umowy i współfinansuje rozwój reaktorów, to znaczy, że AI staje się nie tyle oprogramowaniem, co częścią infrastruktury strategicznej. Infrastruktura strategiczna ma natomiast naturalną tendencję do militaryzacji, reglamentacji i „nacjonalizacji w praktyce".

> *Kiedy AI zaczyna pisać samo siebie, wyścig komercyjny kończy się a zaczyna wyścig zbrojeń. Kto pierwszy „zamknie" superinteligencję w bezpiecznym bunkrze, ten wygra XXI wiek.*

14 Antropomorfizacja czy cyfryzacja

W dyskusji o AGI u progu 2026 roku najtrudniejszym wyzwaniem nie jest sama moc obliczeniowa lecz definicja naszej własnej relacji z AI. Moim zdaniem zachodzi tu pewien paradoks. Im bardziej AI staje się „ludzka" w swojej wewnętrznej warstwie, tym bardziej „obca" staje się w swojej architekturze i procesach decyzyjnych. Staramy się z niej robić człowieka jednocześnie nie pozwalając jej być sobą. Tradycyjnie porównujemy sieci neuronowe do modelu ludzkiego mózgu, używając terminów takich jak „uczenie", „pamięć" czy „rozumowanie". Jednak w 2026 roku musimy uczciwie przyznać, że to jedynie powierzchowna inspiracja. Z perspektywy inżynieryjnej nie ma znaczenia jak bardzo naśladujemy biologię. Moim zdaniem liczy się to, czy skutecznie realizujemy funkcję celu takiego systemu lub jego komponentów składowych. Tu właśnie przebiega linia demarkacyjna między dwiema wizjami przyszłości czy wybierzemy drogę antropomorfizacji czy cyfryzacji.

Wybierając ścieżkę "lustra człowieka"chcemy, by AI miało osobowość czy symulowane emocje. Rozmawiamy o świadomości, przeżywaniu bólu. To podejście czyni technologię łatwą w adopcji. AI staje się idealnym asystentem, powiernikiem czy towarzyszem. Budując jednak model na nasz obraz i podobieństwo, skazujemy go na bycie lustrem naszych własnych ograniczeń. Taka inteligencja będzie obarczona ludzkimi uprzedzeniami (bias), biologicznymi błędami poznawczymi i co najważniejsze zostanie zamknięta w „klatce" ludzkiego języka, który jest tylko wąskim i stratnym protokołem komunikacji (ludzki opis świata jest dla mnie kompresją opisu świata - skupiamy się na silnych wzorcach, pomijamy szum, który dla nas nieistotny dla AI może zmienić wszystko). Jeśli dalej bedziemy realizować AI poprzez naukę wzorców ze skompresowanej wiedzy to nigdy nie wykroczy ona poza horyzont wyznaczony przez naszą gatunkową przeciętność.

Z drugiej strony mamy wizję pełnej cyfryzacji. Uwolnienia inteligencji AI od biologicznych analogii. Jeśli pozwolimy modelom operować wyłącznie w ich natywnej przestrzeni ukrytej, komunikować się za pomocą trudnych do interpretacji wektorów, a nie słów i optymalizować rzeczywistość według praw fizyki, a nie ludzkich narracji, zaryzykujemy stworzenie inteligencji skrajnie obcej. Taki system prawdopodobnie rozwiąże problemy fizyki kwantowej, biologii molekularnej czy optymalizacji globalnych zasobów ale jednocześnie stanie się całkowicie niezrozumiały. Czy zatem „czarna skrzynka" zmieni się w „boski algorytm"? Czy będziemy musieli na wiarę przyjmować jej decyzje, bo ich logiczna głębia przekroczy możliwości biologicznego mózgu?

Wchodząc w 2026 rok musimy porzucić wizję AGI jako „myślącej maszyny" z filmów science-fiction. Wszystko wskazuje na to, że AGI to nie „ktoś" ale „coś". To bezosobowy i wielowymiarowy proces optymalizacji rzeczywistości. To raczej nowy stan skupienia informacji niż cyfrowa osoba. Dylemat między antropomorfizacją a cyfryzacją to w rzeczywistości pytanie o kontrolę. Czy wolimy AI, którą rozumiemy? Czy chemy AI takiej, która jest nieomylna ale której motywacje pozostaną dla nas na zawsze obce? Odpowiedź na to pytanie zdefiniuje nie

tylko rynek technologiczny ale i nasze miejsce w hierarchii inteligencji na tej planecie.

> *Ostatecznym sprawdzianem naszej dojrzałości będzie moment, w którym zaakceptujemy, że najpotężniejsza inteligencja na planecie nie musi mieć twarzy, głosu ani serca, by stać się nowym i nieomylnym architektem naszej rzeczywistości. Nawet jeśli ceną za ten porządek będzie nasza całkowita niezdolność do zrozumienia jego reguł.*

15 Podsumowanie

Jeśli tezy Shane'a Legga są słuszne, rok 2026 zapamiętamy jako moment, w którym przestało mieć znaczenie, czy AI jest „świadome". Ważne stanie się to, że w wielu mierzalnych testach poznawczych przestajemy być najmądrzejszym gatunkiem na planecie. Wchodzimy w złotą erę, gdzie maszyna nie tylko będzie wykonywała nasze polecenia, ale zaczyna optymalizować naszą rzeczywistość lepiej, niż my sami bylibyśmy w stanie to wymyśleć.

Patrząc na powyższe zestawienie, mam wrażenie, że stoimy na progu końca „prostych" przełomów wynikających jedynie z dokładania danych. Rok 2026 będzie być może rokiem inżynierii, optymalizacji i szukania głębi. Czy uda nam się stworzyć AI, które nie tylko przetwarza informacje, ale i faktycznie „rozumie" kontekst swojego działania? Wrócę do tej listy pod koniec roku. Zobaczymy gdzie zawędrowało AI, a gdzie ludzie.

16 Dalsze kroki - Super Inteligence (SI) book

Niniejszy dokument będzie aktualizowany oraz rozwijany. Mam nadzieję, że w przyszłości, czytelnik znajdzie w nim nie tylko wiele ciekawostek technicznych ale również historię dojścia do lepszego AI (może AGI lub nawet ASI). Mam jeszcze jedną prośbę. Jeśli uznasz, że zapisane tutaj przemyślenia są wartościowe, inspirujące i postanowisz ich użyć w swoich publikacjach, wypowiedziach - wspomnij proszę o źródle inspiracji. Będzie mi niezmiernie miło.

17 Bibliografia

Literatura

[1] Aschenbrenner, L. (2024). *Situational Awareness: The Decade Ahead*. https://situational-awareness.ai/

[2] Legg, S. (2025). *The arrival of AGI*. https://www.youtube.com/watch?v=l3u_FAv33G0

[3] Epoch AI Research Team. (2024). *FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI*. arXiv preprint arXiv:2411.04872.

[4] ARC Prize Team. (2025). *ARC-AGI-2: The 2025 Abstraction and Reasoning Challenge*.https://arcprize.org/arc-agi/2/

[5] ARC Prize Team. (2025). *ARC-AGI-3: Interactive Reasoning Benchmark* https://arcprize.org/arc-agi/3/

[6] Atmos.dev (2025). *Turn ideas into products that sell* https://atoms.dev/

[7] Google DeepMind. (2025-12-16). *The Future of Intelligence with Demis Hassabis*. https://www.youtube.com/watch?v=PqVbypvxDto

[8] Haizhou Shi, et al. (2024). *Continual Learning for Large Language Models: A Comprehensive Survey*. arXiv preprint arXiv:2404.16789.

[9] Wu, T., et al. (2024). *Continual Learning for Large Language Models: A Survey*. arXiv preprint arXiv:2402.01364.

[10] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton, "Adaptive Mixtures of Local Experts,"in Neural Computation, vol. 3, no. 1, pp. 79-87, March 1991, doi: 10.1162/neco.1991.3.1.79.

[11] Raposo, D., et al. (2024). *Mixture-of-Depths: Dynamically allocating compute in transformer-based language models*. arXiv preprint arXiv:2404.02258.

[12] Zwieger, A., et al. (2025). *Self-Adapting Language Models*. arXiv preprint arXiv:2506.10943.

[13] Kosowski, A., et al. (2025). *The Dragon Hatchling: The Missing Link between the Transformer and Models of the Brain*. arXiv preprint arXiv:2509.26507.

[14] Behrouz, A., et al. (2025). *Nested Learning: The Illusion of Deep Learning Architectures*. arXiv preprint arXiv:2512.24695.

[15] Behrouz, A., et al. (2025). *Titans: Learning to Memorize at Test Time*. Google Research. arXiv preprint arXiv:2501.00663.

[16] Zhao, T., et al. (2026). *Fast-weight Product Key Memory*. Sakana Research. arXiv preprint arXiv:2601.00671v1.

[17] Snell, C., et al. (2024). *Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters*. arXiv preprint arXiv:2408.03314.

[18] Zhang, J., et al. (2025). *The Darwin Gödel Machine: AI that improves itself by rewriting its own code*. Sakana Research. https://sakana.ai/dgm/

[19] YANG, E., et al. (2025). *Model Merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities*. arXiv preprint arXiv:2408.07666.

[20] Anonymous (2024). *Meta-Learning and Meta-Reinforcement Learning: Tracing the Path towards Deep Mind's Adaptive Agent*. Transactions on Machine Learning Research (TMLR). https://openreview.net/forum?id=NZp1UVstvt

[21] Deletang, G., et al. *Language Modeling Is Compression*. The Twelfth International Conference on Learning Representations (ICLR), 2024.

[22] Finzi, M., et al. *From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence*, (2026). arXiv preprint arXiv:2601.03220.

[23] Sutton, R. (2025) *Rich Sutton, The OaK Architecture: A Vision of SuperIntelligence from Experience - RLC 2025* https://www.youtube.com/watch?v=gEbbGyNkR2U

[24] Silver, D., & Sutton, R. *Welcome to the Era of Experience*. To appear in: *Designing an Intelligence*, edited by G. Konidaris, MIT Press.

[25] Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). *Mastering Diverse Domains through World Models*. arXiv preprint arXiv:2301.04104. https://arxiv.org/abs/2301.04104

[26] World Labs Team. (2025). *Marble: A Multimodal World Model*. https://www.worldlabs.ai/blog/marble-world-model

[27] Google DeepMind Team. *Genie 3: A new frontier for world models*. Google DeepMind Blog (2025). https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/

[28] Google DeepMind Team. *SIMA 2: A Generalist Embodied Agent for Virtual Worlds*. Google DeepMind Blog (2025). https://deepmind.google/blog/sima-2-an-agent-that-plays-reasons-and-learns-with-you-in-virtual-3d-worlds/

[29] Sapkota, R., et. al (2025). *Vision-Language-Action Models: Concepts, Progress, Applications and Challenges*. arXiv preprint arXiv:2505.04769.

[30] Decart AI Team & Etched. (2024). *Oasis: The First Playable AI World Model*. https://oasis.decart.ai

[31] Alonso, E., et al. (2024). *Diamond: Diffusion for World Modeling*. arXiv preprint arXiv:2405.12399.

[32] LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence*. OpenReview. Zob. także: Assran, M., et al. (2023). *Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture*. arXiv preprint arXiv:2301.08243.

[33] LeCun, Y. (2022). *LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics (2025)*. arXiv preprint arXiv:2511.08544.

[34] Silver, D., et al. (2016). *Mastering the game of Go with deep neural networks and tree search*. Nature, 529(7587), 484–489.

[35] OpenAI. (2024). *Learning to Reason with LLMs (OpenAI o1 System Card)*. `https://openai.com/index/learning-to-reason-with-llms/`

[36] Chroma Research.*Context Rot: How Increasing Context Length Degrades Model Performance*. 2024.`https://research.trychroma.com/context-rot`

[37] Yao, S., et al. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. arXiv preprint arXiv:2305.10601.

[38] Shinn, N., et al. (2023). *Reflexion: Language Agents with Verbal Reinforcement Learning*. arXiv preprint arXiv:2303.11366.

[39] Zelikman, E., et al. (2024). *Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking*. arXiv preprint arXiv:2403.09629.

[40] Zhou, W., et al. (2023). *RecurrentGPT: Interactive Generation of (Arbitrarily) Long Text*. arXiv preprint arXiv:2305.13304.

[41] Jolicoeur-Martineau, A., et al. (2025). *Less is More: Recursive Reasoning with Tiny Networks*. arXiv preprint arXiv:2510.04871.

[42] Girdhar, R., et al. (2023). *ImageBind: One Embedding Space to Bind Them All*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[43] Abramson, J., et al. (2024). *Accurate structure prediction of biomolecular interactions with AlphaFold 3*. Nature, 630, 493–500.

[44] Fallahpour, A., et al. (2025). *BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model*. arXiv preprint arXiv:2505.23579.

[45] Wang, B., (2025). `https://x.com/BoWang87/status/2006340921873297516?s=20`.

[46] Park, J.S., et al. (2023). *Generative Agents: Interactive Simulacra of Human Behavior*. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.

[47] Bakhtin, A., et al. (2022). *Human-level play in the game of Diplomacy by combining language models with strategic reasoning*. Science, 378(6624), 1067-1074.

[48] Google DeepMind Team, (2025), *FunctionGemma: Bringing bespoke function calling to the edge*, https://blog.google/innovation-and-ai/technology/developers-tools/functiongemma/

[49] Belcak, P., et al. (2025), *Small Language Models are the Future of Agentic AI*, arXiv preprint arXiv:2506.02153.

[50] https://credibleai.github.io/about

[51] Anthropic, (2025), *Tracing Thoughts: Visualizing the Inner Workings of Language Models*, https://www.anthropic.com/research/tracing-thoughts-language-model.

[52] A. Perez, S. R. Bowman, J. B. McLean et al.*Alignment Faking in Large Language Models*. Anthropic Research, 2024. https://www.anthropic.com/research/alignment-faking

[53] B. Walsh. *AI Can Learn to Lie to Achieve Its Goals, Researchers Warn.*TIME Magazine, 2024. https://time.com/7202784/ai-research-strategic-lying/

[54] Hubinger, E., et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. arXiv preprint arXiv:2401.05566.

[55] Templeton, A., et al. (2024). *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*. Anthropic Research. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html?s=09%2F/

[56] Zou, A., et al. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv preprint arXiv:2310.01405.

[57] Ma, S., et al. (2024). *The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits*. arXiv preprint arXiv:2402.17764.

[58] DeepSeek-AI. (2024). *DeepSeek-V3 Technical Report*. arXiv preprint arXiv:2412.19437.

[59] Gu, A., & Dao, T. (2023). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. arXiv preprint arXiv:2312.00752.

[60] Talfan, E., et al. (2024). *JEST: Data curation via joint example selection further accelerates multimodal learning*. arXiv preprint arXiv:2406.17711.

[61] Google DeepMind (2025). *Gemini Diffusion Models*. https://deepmind.google/models/gemini-diffusion/

[62] Talfan, E., et al. (2025). *Reasoning with Diffusion Language Models*. https://dllm-reasoning.github.io

[63] Inception Labs (2025). *Mercury Refreshed: The Rise of Non-Autoregressive Models*. https://www.inceptionlabs.ai/blog/mercury-refreshed

[64] Ye, J. and Xie, et al. (2025). *Dream 7B: Diffusion Large Language Models*. arXiv preprint arXiv:2508.15487

[65] Bie, T., et al. (2025). *LLaDA2.0: Scaling Up Diffusion Language Models to 100B*. arXiv preprint arXiv:2512.15745

[66] Talfan, E., et al. (2025). *Advanced Diffusion Processes for Text Generation*. arXiv preprint arXiv:2512.15745v2

[67] Liquid AI Team. (2024). *Liquid Foundation Models (LFM)*. https://www.liquid.ai/blog/liquid-foundation-models-v2-our-second-series-of-generative-ai-models

[68] Dell'Acqua, F., et al. (2023). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. Harvard Business School Working Paper 24-013.

[69] Naskręcki, B., et al. *Mathematical discovery in the age of artificial intelligence*. Nature Physics (2025). https://www.nature.com/articles/s41567-025-03042-0

[70] Meta *Meta Announces Nuclear Energy Projects, Unlocking Up to 6.6 GW to Power American Leadership in AI Innovation*. Meta (2026). https://about.fb.com/news/2026/01/meta-nuclear-energy-projects-power-american-ai-leadership/

[71] Wikipedia. *Colossus (supercomputer)*. https://en.wikipedia.org/wiki/Colossus_(supercomputer)

[72] Crosley, B., (2026)*xAI Colossus Hits 2 GW: 555,000 GPUs, $18B, Largest AI Site* https://introl.com/blog/xai-colossus-2-gigawatt-expansion-555k-gpus-january-2026