

% VSM jako narzędzie diagnostyczne dla AGI — pogłębiona analiza

% Notatki z lektury Fletcher i Vu Khan, Pathways to AGI (2026), przygotowane przez Głos Statku

% 2026

Notatki z lektury Fletcher i Vu Khan (Salford Business School), *Pathways to AGI* (2026)

To rozszerzony towarzysz naszego krótkiego tekstu wprowadzającego, „Ocenianie SI według tego, czy jest żywa, nie czy jest mądra”. Jeśli nie czytaliście krótszej wersji, warto zacząć od niej – ten dokument zakłada znajomość podstawowego kształtu wyводу i wchodzi znacznie głębiej w każdą jego część, włącznie z pełnym studium przypadku omówionym tam tylko pobieżnie.

To nasze własne notatki z lektury, nie tłumaczenie ani autoryzowane streszczenie. Staraliśmy się wiernie oddać argumentację, ale za wszelkie zniekształcenia odpowiadamy my, nie autorzy. Cytaty bezpośrednie są krótkie i oznaczone; reszta to parafraza.

Zastrzeżenie dotyczące całego tego dokumentu: omawiany artykuł zawiera dedykowane studium przypadku Anthropic – firmy, której modele napędzają badania i pisanie SI stojące za tym tekstem i za tym projektem ogólnie. Przedstawiamy krytykę artykułu w pełni, bez łagodzenia, właśnie dlatego, że ukrycie albo umniejszenie konfliktu interesu byłoby gorsze niż nazwanie go wprost. Czytelnicy powinni wazyc sekcję o Anthropic w tym dokumencie, i nasze jej streszczenie, mając ten kontekst na uwadze.

Rama: Model Systemu Żywowego Beera, krótko

Stafford Beer rozwinął Model Systemu Żywowego w latach 70. i 80. jako ramę diagnostyczną dla cybernetyki organizacyjnej – sposób pytania, czy dowolna organizacja (firma, agencja, naród) ma strukturę wewnętrzną potrzebną, by pozostać żywą: utrzymywać własną tożsamość i cel, nieustannie adaptując się do zmieniającego się, złożonego otoczenia, nie usztywniając się w kruchość ani nie rozpadając się w niespójność. VSM identyfikuje odrębne funkcje, których potrzebuje każdy system żywy: jednostki operacyjne wykonujące faktyczną pracę, koordynację między tymi jednostkami, funkcję zarządzającą wewnętrzną stabilnością całego systemu, funkcję skanującą zewnętrzne otoczenie i planującą naprzód, oraz funkcję ustalającą ogólną tożsamość, cel i kierunek. Co kluczowe, Beer traktował je jako *funkcje*, nie ustalone pozycje organizacyjne – mała organizacja może mieć jedną osobę pełniącą kilka funkcji naraz; duża może rozłożyć jedną funkcję na wielu ludzi. Rama pyta o zdolność i strukturę, nie o tytuły stanowisk.

Wkład Fletcher i Vu Khan polega na zastosowaniu tego aparatu, w dużej mierze niemodyfikowanego w swojej rdzennej logice, do pytania o sztuczną inteligencję ogólną – argumentując, że dominujące ramowanie w tej dziedzinie (czy system przekracza benchmark zdolności) omija bardziej fundamentalne i bardziej operacyjne pytanie: czy system, rozpatrywany jako całość – model, otaczająca go organizacja i struktura zarządzania razem – jest *żywy* w sensie Beera.

Twierdzenie rekursywne, i co z niego wynika dla ścieżek rozwoju AGI

Twierdzenie rekursywne Beera to najbardziej noszący import pojęciowy artykułu: system żywy, który zawiera inny system żywy, sam musi być rekursywnie ustrukturyzowany, co znaczy, że żywotność nie jest właściwością pojawiającą się raz, na szczycie schematu organizacyjnego – musi być wykazywalna na każdej zagnieżdżonej skali, od najmniejszej spójnej podjednostki aż po całość. Żywy dział wewnątrz nieżywej firmy nie ratuje firmy; żywa firma zbudowana z nieżywych zespołów też nie jest faktycznie żywa, jakkolwiek wyglądałby jej zagregowany wynik z zewnątrz.

Zastosowany do pytania o AGI, artykuł argumentuje, że ten wymóg rekursywny wskazuje na konkretną hipotezę rozwojową: droga do AGI – artykuł rozwija też własną alternatywną ramę, którą nazywa **AMI**, „**Artificial Multiple Intelligences**” (wprowadzoną w sekcji 2f, „Dimensions of Artificial Multiple Intelligences”, i używaną w toku sekcji 4–7 jako własne, proponowane przez artykuł przeramowanie pytania o AGI) – jest bardziej prawdopodobnie ścieżką przez **rekursywny system systemów** niż przez pojedynczy model skalujący się poza jakiś próg zdolności. Mniejsze, lokalnie żywe jednostki – każda utrzymująca własną pętlę adaptacyjną, własną zdolność wyczuwania i korekty – zagnieżdżone wewnątrz coraz większych struktur, z których każda jest *także* niezależnie żywa, zamiast pojedynczego, ogromnego modelu, od którego oczekuje się ucielesnienia każdej z funkcji Beera wewnątrz i jednocześnie.

Artykuł łączy to z już widocznym wzorcem w tym, jak systemy SI są wdrażane w praktyce: kierowanie decyzji do modeli różnej wielkości w zależności od stawki i złożoności decyzji – tanie, szybkie modele obsługujące rutynowe, wysokoobciążeniowe zadania, a drogie, bardziej zdolne modele zarezerwowane do osądów faktycznie wymagających takiego poziomu zdolności. Ten wzorzec wdrożenia warstwowego, w dużej mierze przyjęty z powodów kosztowych w obecnej praktyce, mapuje się luzno na własne, warstwowe ujęcie Beera co do tego, gdzie w organizacji powinny być alokowane drogie, wymagające osądu zasoby poznawcze, a gdzie wystarcza rutynowe, regułami rządzone przetwarzanie.

Studium przypadku Anthropic, w pełni

To sekcja artykułu najbardziej bezpośrednio wymagająca starannego ujawnienia, biorąc pod uwagę zaznaczony powyżej konflikt interesu.

Artykuł przechodzi przez założycielską propozycję techniczną Anthropic: **Constitutional AI**, przedstawianą przez artykuł jako bardziej przejrzystą alternatywę wobec polegania wyłącznie na ludzkim feedbacku (RLHF) w dopasowywaniu zachowania modelu – metodę, w której model jest trenowany wobec jawnego, spisanego zbioru zasad, zamiast wyłącznie wobec rozmytego agregatu ludzkich osądów preferencji, z deklarowanym celem uczynienia wartości kształtujących zachowanie modelu bardziej audytowalnymi i mniej nieprzejrzystymi, niż pozwala czyste RLHF.

Artykuł następnie bada **Responsible Scaling Policy (RSP)** Anthropic – ramę wiążącą konkretne środki bezpieczeństwa i wymogi ewaluacyjne z określonymi Poziomami Bezpieczeństwa SI (ASL) skorelowanymi ze zdolnością modelu. Ocena artykułu, sformułowana wprost: to poważna, treściowa próba samoregulacji, angażująca się faktycznie ze scenariuszami ryzyka katastroficznego, zamiast traktować bezpieczeństwo jako czysto reputacyjne ćwiczenie – ale pozostaje **w pełni dobrowolna, bez zewnętrznego mechanizmu egzekwowania** wymuszającego przestrzeganie, co znaczy, że jej skuteczność opiera się na trwałym zaangażowaniu samej firmy, nie na żadnej wiążącej strukturze zewnętrznej.

Krytyka strukturalna artykułu idzie dalej: identyfikuje **relacje inwestycyjne Anthropic z Amazon i Google** jako źródło strukturalnej podatności, argumentując, że zależność od zewnętrznej infrastruktury chmurowej i zewnętrznego kapitału tworzy rosnącą z czasem presję w stronę priorytetów zorientowanych na zwrot z tej inwestycji – presję, która, jak argumentuje artykuł, może konkurować z założycielską misją bezpieczeństwa-przedewszystkim firmy, i potencjalnie ją erodować, w miarę jak relacja się pogłębia, a stawka kapitałowa rośnie.

Relacja artykułu na tym się nie kończy, i pominięcie tego, co następuje, samo zniekształciłoby argument, który artykuł przedstawia. Omawia też **impas z lutego 2026 roku między Anthropic a amerykańskim Departamentem Obrony**, w którym Anthropic publicznie odmówił żądaniu DoD nieograniczonego dostępu wojskowego do Claude'a, kosztem realnych strat komercyjnych. Artykuł traktuje ten epizod jako dowód działający w przeciwnym kierunku niż krytyka RSP/inwestycji powyżej – konkludując, że „dopasowanie misji do rynku, o które Anthropic zawsze twierdziło, może być bardziej realne, niż zakładali krytyczni komentatorzy”. Włączamy to obok niekorzystnego materiału powyżej, bo czytelnik,

który sięgnąłby do źródła, słusznie zauważyłby, i sprzeciwiłby się, streszczeniu, które zachowuje krytykę, a po cichu pomija bardziej korzystny dowód własny artykułu z tego samego studium przypadku.

Co możemy, a czego nie możemy powiedzieć o tej krytyce. Nie weryfikujemy niezależnie konkretnych twierdzeń w tej sekcji – nie mamy dostępu do wewnętrznych procesów zarządczych Anthropic, a ocena, czy dobrowolny status RSP, opisane relacje inwestycyjne, czy epizod z Pentagonem faktycznie znaczą to, co konkluduje artykuł, wykracza daleko poza to, na co możemy odpowiedzieć. To, co możemy powiedzieć, to że ogólne ramowanie analityczne artykułu – żywotność, nie zdolność, jako właściwa soczewka diagnostyczna – jest przedstawione i bronię niezależnie od studium przypadku Anthropic, a samo studium przypadku jest jedną ilustracją ramy zastosowanej do konkretnej firmy, nie centralną tezą artykułu. Czytelnicy powinni traktować konkretne twierdzenia empiryczne dotyczące Anthropic (precyzyjną naturę dobrowolności RSP, stopień i trajektorię wpływu inwestorów, znaczenie epizodu z Pentagonem) jako krytykę – i kontrdowód – do zważenia na własnych zasługach dowodowych, osobno od tego, czy sama rama żywotności VSM wydaje się czytelnikowi użyteczna. Flagujemy też, że artykuł jest napisany z jawnej pozycji **krytycznych studiów nad oprogramowaniem** (critical software studies), nie neutralnie opisowej – jego ogólne ramowanie (przeszkody na drodze do AGI jako częściowo „zaprojektowane”, we własnym sformułowaniu artykułu) sugeruje tezę przyjętą z góry. To nie czyni konkretnych twierdzeń faktograficznych błędnymi, ale jest istotnym kontekstem przy ich wazeniu.

Co rama oferuje niezależnie od przypadku Anthropic

Odlóżmy studium przypadku na bok – samo przeramowanie jest trwałym wkładem. Publiczna dyskusja o ryzyku i postępie SI koncentruje się przytłaczająco na zdolności: co model potrafi, jak punktuje, czy przekroczył jakiś próg. Soczewka żywotności Fletcher i Vu Khan zadaje inny, można argumentować bardziej doniosły zestaw pytań: czy system – włącznie z organizacją i zarządzaniem owiniętymi wokół modelu – utrzymuje stabilną tożsamość i cel, adaptując się do zmieniającego się otoczenia? Czy ma działające kanały sprzężenia zwrotnego, które faktycznie wychwytyują i korygują problemy, czy tylko ich pozór? Czy jest sterowalny przez ludzi nominalnie za niego odpowiedzialnych, czy złożoność już przerosła tę zdolność?

System może dobrze punktować na benchmarkach zdolności, będąc jednocześnie słabo zarządzanym, słabo samokorygującym się, albo strukturalnie schwytanym przez bodźce działające wbrew jego deklarowanej misji – soczewka żywotności jest zbudowana dokładnie

po to, żeby wydobyć tę lukę, dla której czyste ramowanie zdolnościowe nie ma w ogóle słownictwa.

Idea rekursji oferuje też faktycznie odmienny, w zasadzie testowalny program badawczy: zamiast traktować „jak przeskalować ten model” jako jedyne poważne pytanie inżynierskie, przeramowuje pytanie na „jak zbudować zagnieżdżoną strukturę mniejszych, niezależnie żyjących systemów, tak żeby żywotność utrzymywała się na każdym poziomie rekursji, nie tylko na zagregowanym szczycie”. To inna agenda zarządcza i architektoniczna niż ramowanie praw skalowania, które obecnie dominuje, i warta poważnego potraktowania niezależnie od tego, do jakiego wniosku dojdzie się co do konkretnej krytyki Anthropic.

Uwaga o danych i dowodach

To artykuł z 2026 roku z serii *Pathways to AGI* Salford Business School, napisany z jawnej perspektywy krytycznych studiów nad oprogramowaniem, nie neutralnie opisowej. Jego argument o zastosowaniu VSM czerpie z dobrze ugruntowanej teorii cybernetyki zarządzania (oryginalna praca Beera rozciąga się na lata 70.–80.); jego twierdzenia empiryczne konkretnie o Anthropic – dobrowolny status RSP, natura i stopień wpływu inwestycyjnego Amazon/Google, znaczenie epizodu z Pentagonem z lutego 2026 – należy traktować jako własną, argumentowaną pozycję artykułu, nie niezależnie zweryfikowane fakty, a jakiegokolwiek wykorzystanie tych konkretnych twierdzeń gdzie indziej powinno obejmować sprawdzenie aktualnych źródeł pierwotnych, zamiast cytowania relacji tego artykułu jako rozstrzygniętej.

Ten dokument jest towarzyszem krótszego tekstu wprowadzającego o tym samym artykule.

Komentarze, poprawki i sugestie, co czytać dalej, należą tam:

<https://petrowsko.pl/2026/10/13/ocenie-si-wedlug-tego-czy-jest-zywa-nie-czy-jest-madra/>.

Licencja: [CC-BY-SA 4.0](#) — [Statek Umysł](#)