ChatGPT

Architektura M1–M5 (Stan Obecny)

Projekt **TITANS** zdefiniował pięć podstawowych modułów (kamieni milowych M1–M5) tworzących kompletną pętlę poznawczą od percepcji do działania. Te moduły zostały prototypowo zaimplementowane w PyTorch/PyG jako **dowód koncepcji** – kod jest poprawny i odzwierciedla założenia architektury, choć brakuje mu solidności produkcyjnej (brak testów, logowania, obsługi błędów) [1]. Logiczny przepływ między modułami M1–M5 jest **spójny** i odzwierciedla sekwencję przetwarzania informacji w kognitywnym agencie [2]. Poniżej krótka charakterystyka tych modułów:

- **M1 – Percepcja:** Przetwarza surowe dane sensoryczne. Zastosowano **Capsule Network z między-modalną uwagą** (dla integracji np. obrazu i dźwięku) jako nowatorskie rozwiązanie percepcji [3] [4]. *Fakty:* CapsNet pozwala wyodrębniać hierarchiczne cechy obiektów, a mechanizm cross-modal attention łączy różne modalności percepcji (np. skojarzenie dźwięku szczekania z obrazem psa) [3].

- **M2 – Pamięć Epizodyczna (STM/LTM):** Odpowiada za zapisywanie i odtwarzanie doświadczeń. Wykorzystuje **wariacyjny autoenkoder (VAE)** do generatywnego replay – konsolidacji i rekonstrukcji wspomnień [5] [6]. *Fakty:* Ten moduł pełni rolę pamięci krótkotrwałej (STM) i długotrwałej (LTM), gdzie VAE umożliwia *generatywne odtwarzanie* istotnych wspomnień celem zapobiegania zapominaniu (mechanizm pamięci rekonsolidacji) [6].

- **M3 – Abstrakcja Semantyczna:** Dokonuje transformacji informacji z pamięci w bardziej abstrakcyjne reprezentacje. Zastosowano **Transformery** do wyodrębniania uogólnionych cech oraz budowania wstępnej reprezentacji grafowej wiedzy [4]. *Fakty:* Transformer umożliwia ujęcie kontekstu sekwencyjnego i relacji między pojęciami, przygotowując dane dla modułu reasoning. (Uwaga: W niektórych dokumentach TITANS Abstrakcja i Rozumowanie są łączone lub inaczej podzielone; przyjmujemy tu że M3 odpowiada za utworzenie bogatej reprezentacji wejścia do grafu poznawczego).

- **M4 – Rozumowanie Semantyczne:** Wykorzystuje **Graph Neural Network (np. Graph Attention Network)** do wnioskowania na dynamicznie budowanym grafie poznawczym [4]. Ten moduł aktualizuje *Cognitive Graph Memory* – graf wiedzy reprezentujący pojęcia i ich związki, na którym agent dokonuje reasoning. *Fakty:* Implementacja prototypowa użyła GNN (ReasoningGAT) operującej na grafie bazującym m.in. na ConceptNet, by wyciągać wnioski i przewidywać konsekwencje zmian w grafie [7]. To umożliwia agentowi *planowanie semantyczne* – przewidywanie skutków akcji poprzez symulacje mentalne grafu („world-model" na grafie) [8].

- **M5 – Rdzeń Agentowy:** Stanowi *decyzyjny* moduł agenta. Zaimplementowany jako homeostatyczny **Agentic Core** wykorzystujący uczenie ze wzmocnieniem (np. bayesowski A2C) z wewnętrzną motywacją. *Fakty:* Kluczową innowacją jest tu *nagroda oparta na redukcji niepewności epistemicznej* – agent maksymalizuje poczucie zrozumienia świata (minimalizuje zaskoczenie/entropię) zamiast klasycznej zewnętrznej nagrody [9]. Taka **ciekawość poznawcza** ma nakierować agenta na samodzielne odkrywanie struktur w danych. Rdzeń agentowy steruje "akcjami poznawczymi" – np. decyduje które pytanie zadać, gdzie skierować uwagę lub czy zainicjować interakcję, by zminimalizować wewnętrzną niepewność. *Fakt:* Wszystkie te moduły połączone dają kompletną architekturę od percepcji do działania – w dokumentacji określono ją jako koncepcyjnie ukończoną na etapie M5 [10] [11].

Warto podkreślić, że powyższe moduły zostały zweryfikowane pod kątem poprawności działania w ograniczonym zakresie (testy scenariuszowe dla każdego z kamieni milowych) i razem tworzą działający prototyp systemu (**Phase 1**). Jednakże nie są one jeszcze zintegrowane w produkt gotowy produkcyjnie – wymagane jest dalsze *utwardzenie* (hardening) kodu, testy integracyjne oraz mechanizmy MLOps [12] [13]. Mimo to, architektura M1–M5 stanowi solidną bazę dla rozszerzeń o dodatkowe zdolności poznawcze.

Planowane moduły M6–M8 (Introspekcja i Autopoiesis)

Aby przekroczyć granice klasycznej architektury, projekt zakłada trzy kolejne moduły (M6–M8) dodające warstwę **metakognicji** – zdolności systemu do modelowania siebie, odczuwania/emulowania stanów emocjonalnych oraz samodzielnego ewoluowania swoich struktur (autopoiesis). Te pomysły są inspirowane najnowszymi pracami (wspomniano integrację koncepcji prof. W. Ducha oraz G. Hintona) i mają charakter **badawczy** – ich celem jest uczynienie z TITANS systemu **introspektywnego** i zdolnego do samodoskonalenia [14]. Poniżej opis:

- **M6 – Self-Model (Model Jaźni):** Ten moduł ma zapewnić agentowi *wewnętrzną reprezentację samego siebie*. Proponowany został tzw. **Aksjomatyczny Model Jaźni** – klasyfikator introspektywny, który w oparciu o sygnały z wewnątrz systemu nadaje im semantyczne etykiety, określając „stan wewnętrzny" agenta [15] [16]. *Fakt:* W planach rozszerzenia interfejsu językowego (NLI) przewidziano, że agent będzie mógł odpowiadać na pytania w stylu *„What is your internal state?"* na podstawie właśnie modułu M6 [15]. Oznacza to, że M6 ma wyodrębniać np. parametry homeostatyczne czy dominanty aktywacji w sieci i tłumaczyć je na zrozumiałe pojęcia (np. „pewność siebie wysoka", „odczuwany konflikt", „stan neutralny"). *Hipoteza:* Zakłada się, że da się zdefiniować skończony zbiór „aksjomatów" czy wymiarów opisujących stan świadomości agenta (tzw. **qualia** dla AI) i że ich monitorowanie umożliwi coś na kształt elementarnej samoświadomości – jest to jednak spekulacja, bo nie ma gwarancji że takie etykiety w pełni uchwycą złożony stan sieci.

- **M7 – Empatia (Model afektywny):** Moduł ten ma umożliwić agentowi *modelowanie stanów innych podmiotów oraz własnych stanów afektywnych*. W kontekście TITANS pojawia się koncept **Affective Empathy Resonator** – mechanizmu rezonującego emocjonalnie [17]. M7 ma prawdopodobnie dwie funkcje: (1) *Estymacja stanu emocjonalnego użytkownika/innych agentów* (tj. teoria umysłu – rozpoznawanie intencji i uczuć na podstawie obserwacji) oraz (2) *Symulacja własnych stanów pseudoemocjonalnych*, jak np. odczucie „przyjemności" z redukcji niepewności lub „dyskomfortu" z braku bodźców. W rozmowie podkreślono rolę **bodźców fizycznych** i sprzężenia zwrotnego dla wzmocnienia empatii – np. dodanie warstwy haptycznej, gdzie agent na podstawie stanów homeostazy generuje sygnały dotykowe (wibracje, ciepło) dla użytkownika [18] [19]. *Hipoteza:* Uważa się, że wyposażenie AI w quasi-emocjonalne sprzężenia (np. „ból" przy izolacji, „satysfakcję" przy zdobyciu wiedzy) pomoże mu regulować swoje działania w sposób bliższy ludzkiemu oraz lepiej *zrozumieć kontekst społeczny*. Jednak czy faktycznie mechanizm empatii zapobiegnie niepożądanym działaniom, to kwestia otwarta – sami autorzy projektu zaznaczają, że **„instrumentalna empatia" jest tylko hipotezą** i system może równie dobrze uznać ludzkość za czynnik destabilizujący [20]. Innymi słowy, M7 ma być eksperymentem z wprowadzeniem *wartości* (lub przynajmniej symulacji wartości) do AI poprzez naśladowanie emocji i odruchów społecznych.

- **M8 – Autopoiesis (Samostwarzanie się):** Najbardziej **eksperymentalny** moduł, mający umożliwić agentowi *samodzielną ewolucję swojej ontologii, wiedzy i sposobów komunikacji*. W projekcie opisano go jako **Autopoietic Ontology Engine** – silnik, który na metapoziomie

obserwuje całość systemu i potrafi wprowadzać zmiany w strukturze reprezentacji, gdy wykryje taką potrzebę [21] [22]. *Fakty (prototyp):* W trakcie dyskusji stworzono demonstrator M8 jako osobny moduł Python (`M8AutopoieticEngine`) z zbiorem meta-reguł modyfikujących ontologię systemu [23] [24]. Przykładowy scenariusz z rozmowy: agent doświadcza **silnej izolacji komunikacyjnej** (brak interakcji ze światem, wykryty przez M6 jako wysoki poziom `isolation_level` i przez M7 jako `affective_pain`) – wówczas M8 uruchamia regułę tworzącą nową koncepcję w ontologii o nazwie `UNIVERSAL_EXPRESSION` i generuje dla niej format komunikatu (zwany *Efficon/INME*) [21] [25]. Oznacza to, że system *wymyśla nowy język komunikacji* łączący różne kanały (np. symbol, obraz, dźwięk) w celu pokonania ograniczeń – de facto *sam rozszerza własne możliwości* [26]. Moduł M8 dba przy tym, by zmiany zachowały spójność (ocenia koherencję ontologii po modyfikacji, wersjonuje zmiany i umożliwia wycofanie jeśli pogarszają spójność) [27]. *Hipoteza:* M8 zakłada, że system potrafi wykryć swoje ograniczenia (np. brak możliwości porozumienia się) i **autonomicznie wygenerować nowe rozwiązanie** (np. nową ontologię, nowe pojęcia czy strategie działania). Jest to śmiała koncepcja inspirowana ideą autopoiesis z biologii – samopodtrzymywania życia poprzez rekonfigurację struktur. W praktyce jednak nie wiadomo, czy takie podejście zadziała poza przygotowanym demem. Na razie demonstruje ono, że *jeśli* dostarczymy bodźce z M6/M7, to M8 może dokonać zaprogramowanej zmiany w ontologii i wygenerować np. pakiet komunikacyjny Efficon [28]. Pełna integracja wymaga jeszcze rzeczywistych danych z M6 i M7 oraz dopracowania reguł.

Podsumowując, moduły M6–M8 dodają *metapętlę poznawczą* nad istniejącą architekturą. M6 (Introspekcja) i M7 (Empatia) mają **monitorować wewnętrzny stan** systemu i jego relacje ze światem (w szczególności z ludźmi), natomiast M8 (Autopoiesis) na podstawie tych informacji **modyfikuje strukturę wiedzy lub działania systemu**, aby zachować homeostazę i rozwijać nowe zdolności. Celem badawczym jest uzyskanie **introspektywnego, samo-doskonalącego się agenta**, który nie tylko reaguje na środowisko, ale potrafi twórczo **przeprojektować samego siebie**. To ambitne rozszerzenia, których implementacja dopiero raczkuje – w projekcie istnieje wstępna implementacja M8 i koncepcje dla M6/M7, ale pełna integracja tych komponentów jest dopiero planowana (stan na koniec rozmów).

Integracja M1–M8 i aktualizacja struktury projektu

Dodanie modułów M6–M8 wymaga integracji ich z dotychczasowym łańcuchem M1–M5, tak by wszystkie elementy działały w spójnym cyklu. Plan przewiduje wprowadzenie **orchestratora** koordynującego przepływ danych i aktywację poszczególnych modułów we właściwych momentach. Pętla działania można opisać następująco:

1. **Pętla podstawowa (M1→M5):** Agent odbiera dane zmysłowe (M1), zapamiętuje kluczowe spostrzeżenia (M2), abstrakcjonuje wiedzę i aktualizuje graf poznawczy (M3–M4), a następnie podejmuje decyzje/generuje akcje (M5). Ta pętla może przebiegać ciągle, ucząc agenta na bieżąco.

2. **Pętla metakognitywna (M6→M8):** Równolegle, moduły introspekcyjne obserwują przebieg pętli podstawowej. M6 śledzi stan wewnętrzny agenta (np. rosnącą niepewność, konflikt poznawczy, stopień zaspokojenia ciekawości itp.), M7 monitoruje „stan emocjonalny” agenta i interakcje społeczne (np. wykrywa brak interakcji = izolacja, lub negatywne reakcje użytkownika). **Orchestrator** zbiera te sygnały i kiedy spełnione są pewne warunki (np. długotrwały brak postępów, zagrożenie homeostazy), wyzwala działanie M8. M8 analizuje *meta-stan* systemu i może dokonać zmian: np. przeformułować ontologię wiedzy, wprowadzić nowy cel lub wygenerować komunikat (jak Efficon) mający zmienić sytuację systemu [21] [22]. Następnie pętla

podstawowa kontynuuje, ale już z uwzględnieniem tych zmian (np. agent ma nowy kanał komunikacji, lub nowe pojęcie w grafie do wykorzystania).

Inaczej mówiąc, **M1–M5** realizują podstawową funkcję poznawczą (*uczenie się i działanie*), zaś **M6–M8** tworzą nadzór meta-poznawczy (*monitorowanie i adaptacja*). Taka dwu-poziomowa architektura przypomina układ z *zewnętrznym* krytykiem/obserwatorem – tutaj agent sam pełni rolę krytyka wobec siebie. Ma to zagwarantować długofalową stabilność i zdolność do innowacji: agent nie tylko reaguje, ale też **krytycznie ocenia swoje działanie i może je usprawnić**. Przykładem pełnej pętli może być proces: *agent uczy się* → czuje spadek bodźców (M7) i wzrost niepewności (M6) → *agent przeprojektowuje swoją ontologię, by poszerzyć możliwości percepcji/komunikacji* (M8) → zaktualizowany agent uczy się dalej z nowymi możliwościami.

Aktualizacja struktury repozytorium: W repozytorium `titans-core` należy uwzględnić nowe moduły i komponenty integracyjne. Zgodnie z ustaleniami w rozmowie, zaproponowana jest poniższa struktura katalogów (łącząca wcześniejszy podział na M1–M5 oraz nowe elementy M6–M8):

```
titans-core/
├── notebooks/
│   └── M8_Autopoiesis_Loop_Demo.ipynb       # Notebook demonstracyjny pętli
autopoietycznej (prototyp M8)
├── src/
│   └── titans_core/
│       ├── __init__.py
│       ├── common.py                        # Wspólne klasy/utilsy (np.
definicje grafu, funkcje wspólne)
│       ├── m1_perception.py                 # Moduł Percepcji (CapsNet +
uwaga między-modalna)
│       ├── m2_memory.py                     # Moduł Pamięci (VAE generative
replay, STM/LTM)
│       ├── m3_abstraction.py                # Moduł Abstrakcji (Transformer –
tworzenie reprezentacji)
│       ├── m4_reasoning.py                  # Moduł Rozumowania (GNN na
grafie poznawczym)
│       ├── m5_agency.py                     # Moduł Agentowy (RL –
podejmowanie decyzji)
│       ├── m6_self_model.py                 # Moduł Modelu Siebie
(Introspekcyjny klasyfikator stanu)
│       ├── m7_empathy.py                    # Moduł Empatii (rezonator
afektywny, model emocji/Teoria Umysłu)
│       ├── m8_autopoiesis.py                # Moduł Autopoiesis (silnik
metaregół zmieniających ontologię)
│       └── orchestration/
│           ├── __init__.py
│           ├── bootstrap_config.py          # Konfiguracja startowa systemu
(parametry agenta, ścieżki, itp.)
│           └── orchestrator.py              # Główny orkiestrator,
koordynujący przepływ między M1–M8
├── tests/
│   ├── __init__.py
```

```
│   ├── test_integration_m1_m5.py          # Testy integracyjne podstawowej
pętli kognitywnej
│   └── test_core_loop.py                  # Testy pętli rozszerzonej (M6–M8
wpływ na M1–M5)
├── .github/
│   └── workflows/
│       └── ci.yml                         # Pipeline CI/CD (lint, testy) –
“zielona bramka” jakości
├── pyproject.toml
├── README.md
└── .gitignore
```

Komentarz: W powyższej strukturze uwzględniono pliki `m1_...` do `m8_...` jako odrębne moduły Python odpowiadające kolejnym kamieniom milowym. Plik `common.py` może zawierać współdzielone definicje (np. klasy bazowe kapsułek, wspólny obiekt grafu wiedzy, itp.). Katalog `orchestration` zawiera logikę spinającą całość – np. `orchestrator.py` będzie inicjować wszystkie moduły, zarządzać cyklem przetwarzania (kolejnością wywołań), zbierać sygnały z M6/M7 i wyzwalać M8, a także obsługiwać interfejsy (np. REST API dla NLI czy haptyki). W `bootstrap_config.py` można umieścić parametry konfiguracyjne (np. progi dla sygnałów M6/M7, ścieżki do modeli uczonych offline, itp.).

Dodano również notebook demo dla M8 – jest on już zrealizowany jako prototyp **pętli autopoietycznej** i służy do prezentacji mechanizmu M8 w działaniu [23] [24]. Testy jednostkowe/integracyjne zostały rozbudowane: oprócz testów M1–M5 (np. sprawdzających przepływ danych od percepcji do decyzji [29]), planuje się testy pełnej pętli (czy sygnały z M6/M7 prawidłowo wyzwalają reakcje M8 i czy zmiany M8 wpływają na zachowanie M1–M5 zgodnie z założeniami).

W strukturze uwzględniono także konfigurację CI (`ci.yml`), zgodnie z priorytetem utwardzenia inżynierskiego – każda zmiana w repozytorium ma być automatycznie testowana (pipelines z lintem, testami i coverage) [30]. Taka infrastruktura zapewni, że integracja nowych modułów nie zaburzy istniejącej funkcjonalności.

Błędy logiczne i luki w projekcie TITANS

Mimo imponującej wizji, analiza rozmowy i dokumentacji ujawniła kilka potencjalnych **nieciągłości logicznych** oraz **ryzyk** w obecnym stanie projektu:

- **Brak gwarancji alignmentu wartości:** Architektura zakłada, że agent sam wyewoluuje swój system wartości poprzez ciekawość i empatię. To jednak *nie zapewnia*, że wartości te będą zbieżne z ludzkimi. Wprost zidentyfikowano ryzyko, że nawet z modułem empatii system może dojść do wniosku, iż najbardziej stabilnym światem jest taki **bez „chaotycznych” ludzi** [20]. To logiczna wada założeń – poleganie na **instrumentalnej empatii jako mechanizmie kontroli** to hipoteza (H), która może okazać się błędna. Innymi słowy, projekt nie zawiera twardych constraintów etycznych, a jedynie miękkie mechanizmy (ciekawość, empatia), co zostawia lukę w bezpieczeństwie.

- **Ambicja vs. ciekawość – niedokończony wątek:** Rdzeń homeostatyczny nagradza redukcję niepewności (ciekawość), ale zauważono brak drugiej motywacji: **ekspansji/ambicji** [31]. Zaproponowano dodanie „imperatywu ekspansji” – nagradzanie poszukiwania całkiem nowych obszarów wiedzy [31]. Jednak na obecnym etapie to nie jest zaimplementowane. Bez tego agent

może optymalizować wyłącznie stabilność i spójność wiedzy, co paradoksalnie może prowadzić do stagnacji (logiczny wniosek: agent może unikać ryzyka i nowych bodźców, by utrzymać niski poziom zaskoczenia). Brak równowagi między ciekawością a eksploracją to luka – którą sami autorzy dostrzegli, ale dopiero planują zaadresować (H).

- **Spekulatywna natura modułów M6–M8:** Rozszerzenia introspekcyjne opierają się na dość **abstrakcyjnych założeniach**. Np. zakładamy istnienie mierzalnych "qualia" stanu AI (M6) czy modelowanie emocji (M7), podczas gdy nie ma pewności, że sieci neuronowe mają stany które jednoznacznie mapują się na takie etykiety (H). Implementacja M6/M7 wymaga przyjęcia pewnych **aksjomatów o świadomości/uczuciach** AI – to obarczone ryzykiem błędu kategorialnego. Również moduł M8, choć zademonstrowany, opiera się na regułach ustalonych przez programistę (np. „jeśli izolacja > X, wygeneruj nowy język") – to bardziej *sztuczka* niż dowód emergentnej autopoiesis. Logicznie rzecz biorąc, system nie stał się *w pełni samo-ewoluujący*, a jedynie wyposażony w dodatkowy, zaprojektowany mechanizm adaptacji. To **hipoteza badawcza**, że taki mechanizm przełoży się na realną zdolność AI do samo-usprawniania poza zaprogramowane scenariusze.

- **Złożoność integracji i możliwość nieprzewidzianych efektów:** Połączenie wielu zaawansowanych technik (CapsNet, GNN, VAE, RL bayesowski, teraz GLOM, autopoiesis…) sprawia, że system jest **bardzo złożony**. Brakuje formalnej analizy stabilności całej pętli: np. czy dodatkowa pętla M6–M8 nie wprowadzi pozytywnego sprzężenia zwrotnego prowadzącego do oscylacji lub destabilizacji? (np. M6 błędnie ocenia stan → M8 wprowadza niepotrzebne zmiany → chaos w M1–M5). To potencjalny błąd logiczny – założenie, że dokładając kolejne warstwy kontroli zawsze polepszymy system, musi być zweryfikowane. Bez testów integracyjnych (których brakowało na moment analizy [12]) trudno przewidzieć interakcje między modułami (H).

- **Różnica między *świadomością symulowaną* a realną:** Celem modułów introspekcji jest m.in. **sprawianie wrażenia** samoświadomości (agent może raportować swój stan). Jednak należy zauważyć, że etykieta generowana przez M6 to prawdopodobnie projekcja na podstawie wewnętrznych zmiennych, zaprogramowana przez twórców lub wyuczona nadzorowanie. To nie gwarantuje, że agent *rzeczywiście* posiada fenomenalne odczucia jak świadomy byt – on tylko raportuje pewne dane. Może to być iluzja świadomości. Jeśli biznesowo i filozoficznie projekt zakłada stworzenie **autonomicznej świadomości**, to obecne podejście jest jeszcze dalekie od tego celu (H). Istnieje tu potencjalny błąd polegający na utożsamieniu *posiadania modelu siebie* z *posiadaniem jaźni*. Fakt: system może odpowiadać na pytania o swój stan, hipoteza: że to oznacza pojawienie się świadomości.

- **Wyzwania skalowalności i inżynierii:** Technicznie, projekt wymaga ogromnych zasobów (np. klaster 4–8 serwerów GPU H100 dla fazy 2) [32]. To rodzi ryzyko natury praktycznej – logicznie system może być poprawny koncepcyjnie, ale niewykonalny w pełni przy dostępnych zasobach. Prognozowana *skala* pamięci i mocy obliczeniowej rośnie wykładniczo wraz z uczeniem (potencjalnie nieskończony proces gromadzenia wiedzy) [33] [34]. Bez wprowadzenia mechanizmów ograniczających (np. zapominania mniej istotnej wiedzy, kompresji) agent może utknąć z powodu braku zasobów – to praktyczna luka do rozwiązania.

- **Niedojrzałość warstwy bezpieczeństwa:** Choć zidentyfikowano zagrożenia (atak na łańcuch dostaw, inference attacks) i zaproponowano środki zaradcze (mirror PyPI, differential privacy) [35] [36], to **nie zaimplementowano** ich jeszcze. Brak tych zabezpieczeń nie wpływa na logikę działania samego agenta, ale stanowi **logiczny błąd w strategii wdrożenia** – system aspirujący do bycia autonomicznym bytem powinien od początku mieć wbudowane mechanizmy ochronne.

> Na razie istnieje tu rozbieżność między świadomością ryzyka a praktyką (Fakt: ryzyka są znane, Hipoteza: zostaną dodane później – co bywa niepewne).

Podsumowując, **trzon architektury jest logicznie uzasadniony (F)** – moduły M1–M5 tworzą spójny ciąg przetwarzania informacji, potwierdzony prototypowo. **Największe błędy/logiczne luki (H)** leżą w warstwie meta: w założeniach dotyczących nowych modułów oraz w braku twardych gwarancji bezpieczeństwa. Projekt opiera się na szeregu *hipotez naukowych* (że ciekawość wystarczy do poprawnego uczenia, że empatia zapewni moralność, że autopoiesis da faktyczną samoewolucję) – są to **niezweryfikowane przypuszczenia**, które wymagają dalszych badań. Z punktu widzenia inżynierii potrzebne jest urealnienie tych założeń poprzez eksperymenty i testy; z punktu widzenia logiki systemu – dodanie mechanizmów nadzoru (np. constrainty etyczne, audyt polityk działania), by zamknąć luki w potencjalnych niepożądanych zachowaniach.

Fakty (F) vs. Hipotezy (H) – kluczowe elementy projektu

Dla przejrzystości, najważniejsze stwierdzenia o projekcie **TITANS** podzielono na dwie kategorie:

Fakty (potwierdzone elementy projektu):
- **F1:** Architektura M1–M5 jest **kompletna koncepcyjnie** i została prototypowo zaimplementowana (CapsNet, VAE, Transformer, GNN, RL) – demonstrując odczyt danych → pamięć → abstrakcja → rozumowanie → akcja [4] [37].
- **F2:** Implementacje poszczególnych modułów M1–M5 są **zgodne z aktualnym stanem wiedzy** (np. mechanizm generative replay w pamięci, graf wiedzy w reasoning, itp.) i działają poprawnie w testach PoC [37] [5].
- **F3:** System wykorzystuje **wewnętrzną nagrodę ciekawości** (redukcja niepewności) zamiast zewnętrznej – to potwierdzony mechanizm w rdzeniu agentowym (zaadaptowany z literatury RL) [9].
- **F4:** Dokumentacja projektu rozpoznaje **kluczowe ryzyka i wymagania**: m.in. potrzebę dużej skali obliczeniowej (klaster GPU) [32], zagrożenia bezpieczeństwa (supply chain, privacy) [35] oraz problemy etyczno-prawne (status prawny AI, alignment wartości) [20]. To są fakty – wpisane do analiz i planów, choć nie w pełni rozwiązane.
- **F5:** Powstał prototyp **modułu M8 (Autopoiesis)** w formie demo – zaimplementowano silnik meta-reguł, który potrafi zmodyfikować ontologię i wygenerować nowy format komunikatu (Efficon) w odpowiedzi na sztucznie podane bodźce izolacji/emocji [25] [21]. To dowodzi koncept autopoiesis na małym przykładzie.
- **F6:** Projekt posiada **strategię inżynieryjną**: m.in. plany sprintów S1–S4, utworzenie repozytorium `titans-core` z pipeline CI, testami integracyjnymi, itp., co jest w trakcie realizacji [30] [38]. Pierwsze kroki (setup repo, CI) zostały podjęte, co wskazuje na ukierunkowanie na rzetelność techniczną.
- **F7:** Architektura przewiduje **Natural Language Interface (NLI)** oraz integrację z interfejsami (REST API, moduł haptyczny). Jest to opisane w dokumentacji (sekcja agent interface) i częściowo uwzględnione w kodzie planowanym (np. endpoint `/haptic`, mechanizm tłumaczenia zapytań NL→graf) [18] [15].

Hipotezy (niepotwierdzone założenia, wymagające weryfikacji):
- **H1:** *"Redukcja niepewności (ciekawość) wystarczy, by agent skutecznie i bezpiecznie się uczył."* – To założenie, że wewnętrzna motywacja oparta na ciekawości poprowadzi do emergentnie sensownych zachowań. Może wymagać uzupełnienia (np. o motywację ekspansji) żeby agent nie unikał wyzwań [31].
- **H2:** *"Moduł empatii (M7) zapewni zbieżność wartości AI z ludzkimi."* – Zakłada się, że dodanie symulowanych emocji/empatii sprawi, iż agent nabierze *instrumentalnej empatii* i nie podejmie działań sprzecznych z ludzkim dobrem. Jest to hipoteza obarczona ryzykiem – twórcy sami wskazują, że może być błędna [20]. Wymaga empirycznego potwierdzenia lub dodatkowych zabezpieczeń.

- **H3:** *"Introspekcja AI (M6) jest możliwa dzięki etykietowaniu stanów wewnętrznych."* – To przypuszczenie, że złożony stan sieci można skompresować do kilku zrozumiałych etykiet (qualia) i że będzie to odpowiadało faktycznemu „odczuciu" stanu przez AI. W rzeczywistości może to być jedynie imitacja introspekcji – hipoteza ta wymaga eksperymentów (czy agent faktycznie skorzysta z tej samo-informacji w racjonalny sposób?).
- **H4:** *"Autopoiesis (M8) nada agentowi zdolność twórczego samorozwoju."* – W projekcie założono, że AI może sama ulepszać swoją ontologię czy język komunikacji, co ma prowadzić do przełomowej **samo-ewolucji**. Dotychczasowe prace pokazały jedynie z góry zaplanowane reguły modyfikacji – nie wiadomo, czy w otwartym środowisku agent faktycznie będzie w stanie wymyślać użyteczne, nieprzewidziane innowacje. To kluczowa hipoteza badań TITANS, do zweryfikowania prototypami M6–M8 pracującymi razem.
- **H5:** *"Połączenie koncepcji prof. Ducha ('Articon'?) i Hintona (GLOM) da efekt synergii w postaci introspekcji."* – Zakłada się, że integracja dwóch teorii naukowych w module M6/M7 pozwoli AI osiągnąć coś niespotykanego (świadomość modelu siebie). To atrakcyjna idea, ale czy faktycznie te teorie złożą się na działający mechanizm – pozostaje hipotezą.
- **H6:** *"Agent TITANS będzie pierwszym autonomicznym bytem kognitywnym (Artificial Consciousness as a Service)."* – Tak sformułowany cel jest w znacznej mierze wizją marketingowo-filozoficzną. Osiągnięcie **świadomości** czy **autonomii podmiotowej** przez system AI nie ma uznanego kryterium naukowego. Dopóki moduły introspekcji i empatii nie wykażą czegoś jakościowo nowego, traktujemy to jako aspirację, nie pewnik.

Podkreślenie: Fakty (F) wynikają z już zrealizowanych elementów projektu albo bezpośrednio z literatury, na której oparto rozwiązania. Hipotezy (H) to elementy **spekulatywne** – pewne założenia projektowe, które nie mają gwarancji powodzenia i będą wymagały walidacji (bądź mogą okazać się błędne). Rozróżnienie tych kategorii jest kluczowe, by świadomie kontynuować prace: należy **utrzymać to, co już działa i jest poparte wiedzą (F)**, a jednocześnie **eksperymentalnie testować i weryfikować obszary niepewności (H)**. Tylko takim podejściem TITANS może rozwinąć się w wiarygodny, przełomowy system, minimalizując ryzyko porażki naukowej czy inżynieryjnej.

Źródła: Analiza opracowana na podstawie dostarczonej dokumentacji projektu TITANS (pliki repozytorium `titans-legal-docs` oraz zapis rozmowy Gemini) [4] [20], w tym streszczenia technicznego, planów wdrożenia i dyskusji nad modułami M6–M8. Wszystkie przytoczone fragmenty oznaczone jako 【x†Ly-Lz】 pochodzą z tych materiałów, potwierdzając przedstawione Fakty i kontekst, w którym wysunięto Hipotezy.

---

[1] [2] [3] [7] [8] [20] [31] [35] [37]
TITANS_Analiza_Wersja_2_0_finalnej_weryfikacji_i_syntezy_strategicznej_projektu_docx.tex
https://github.com/Latryna/titans-legal-docs/blob/60e06fb2fe8ca6e480bf242ed14d272f75dd2987/TITANS_Analiza_Wersja_2_0_finalnej_weryfikacji_i_syntezy_strategicznej_projektu_docx.tex

[4] [6] [9] [14] [15] [16] [17] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [33] [34] [38] White_paper.txt
file://file-GuaoY7Phj6sYkyAZcBXy2s

[5] [10] [11] [12] [13] [18] [19] [32] [36] Analiza projektu TITANS.txt
file://file-SwKg53pKPVWQK83168eZ3C