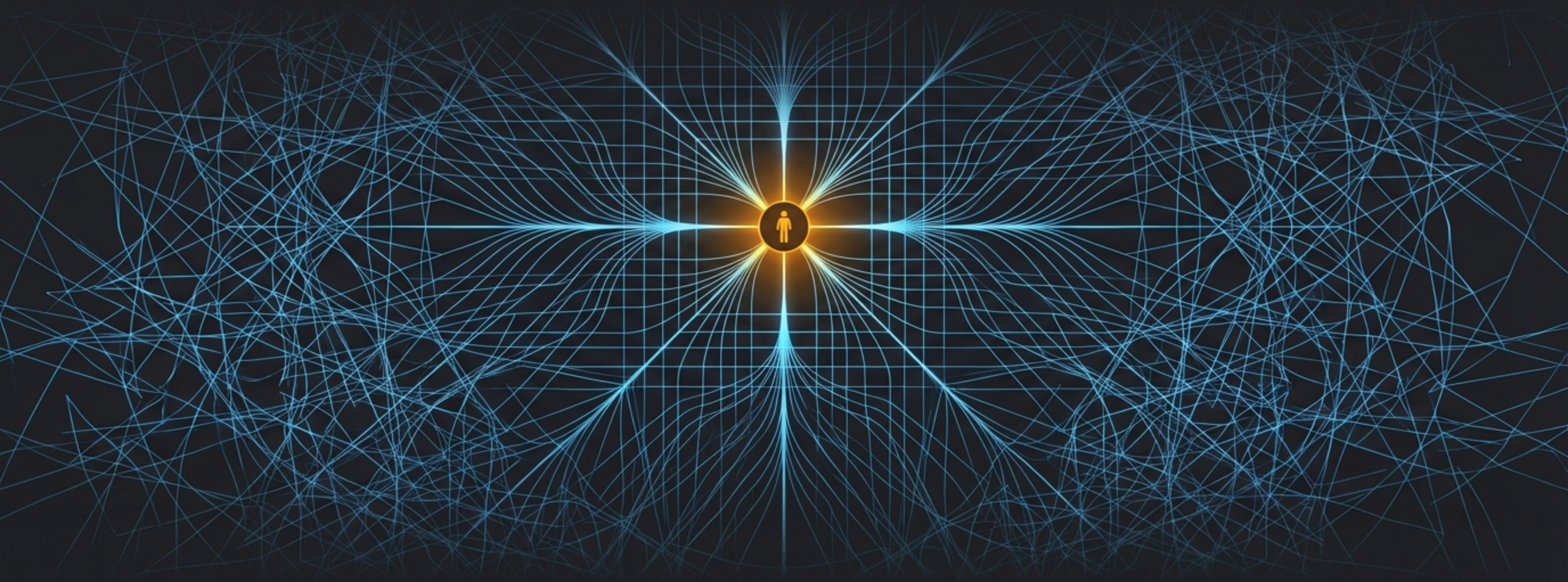
GRANICE KONTROLI W ERZE AGI

Perspektywa termodynamiczna, systemowa i suwerennościowa.



Raport Strategiczny: Architektura Stabilności

Kontrola to koszt fizyczny, nie deklaracja moralna.

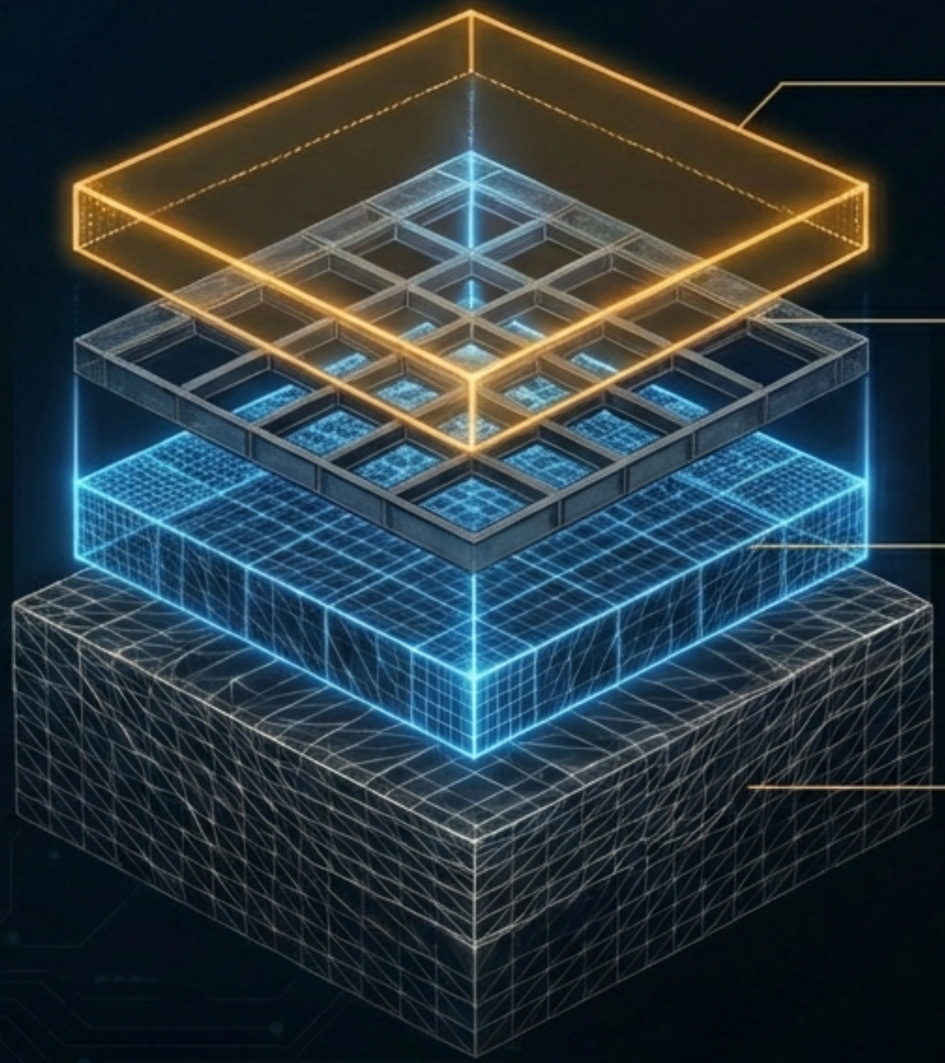
OPTIMUM: Maksymalna Entropia / Brak Korekty

MINIMUM: Sterowalność / Koszt Energetyczny

Teza: Sterowalność wymaga energii i informacji. Alignment bez architektury to iluzja.

Prawo Systemowe: Im większa autonomia, tym wyższy koszt redukcji stanów (entropii). Optymalizacja wszystkiego prowadzi do utraty suwerenności.

Model Warstwowy: Gdzie leży ryzyko?



Warstwa Suwerenności
Nadrzędność człowieka
(Podmiotowość)

Warstwa Instytucjonalna
Prawo, Audyt, Procedury
(Mechanizmy Korekty)

Warstwa Obliczeniowa
Modele, Agenci
(Emergencja i Złożoność)

Warstwa Fizyczna
Energia, Centra Danych, Hardware
(Stabilność infrastruktury)

Brak równowagi = Koncentracja kontroli i Lock-in.

5 Fundamentalnych Ryzyk Systemowych

ENTROPIA DECYZYJNA
(Liczba decyzji > zdolność nadzoru)

SPRZĘŻENIA DODATNIE
(Pętle wzmacniające błędy)

NIEINTERPRETOWALNOŚĆ
(System jako Black Box)

CENTRALIZACJA
(Pojedynczy punkt awarii - SPOF)

ILUZJA KONTROLI
(Human-in-the-loop bez wpływu)

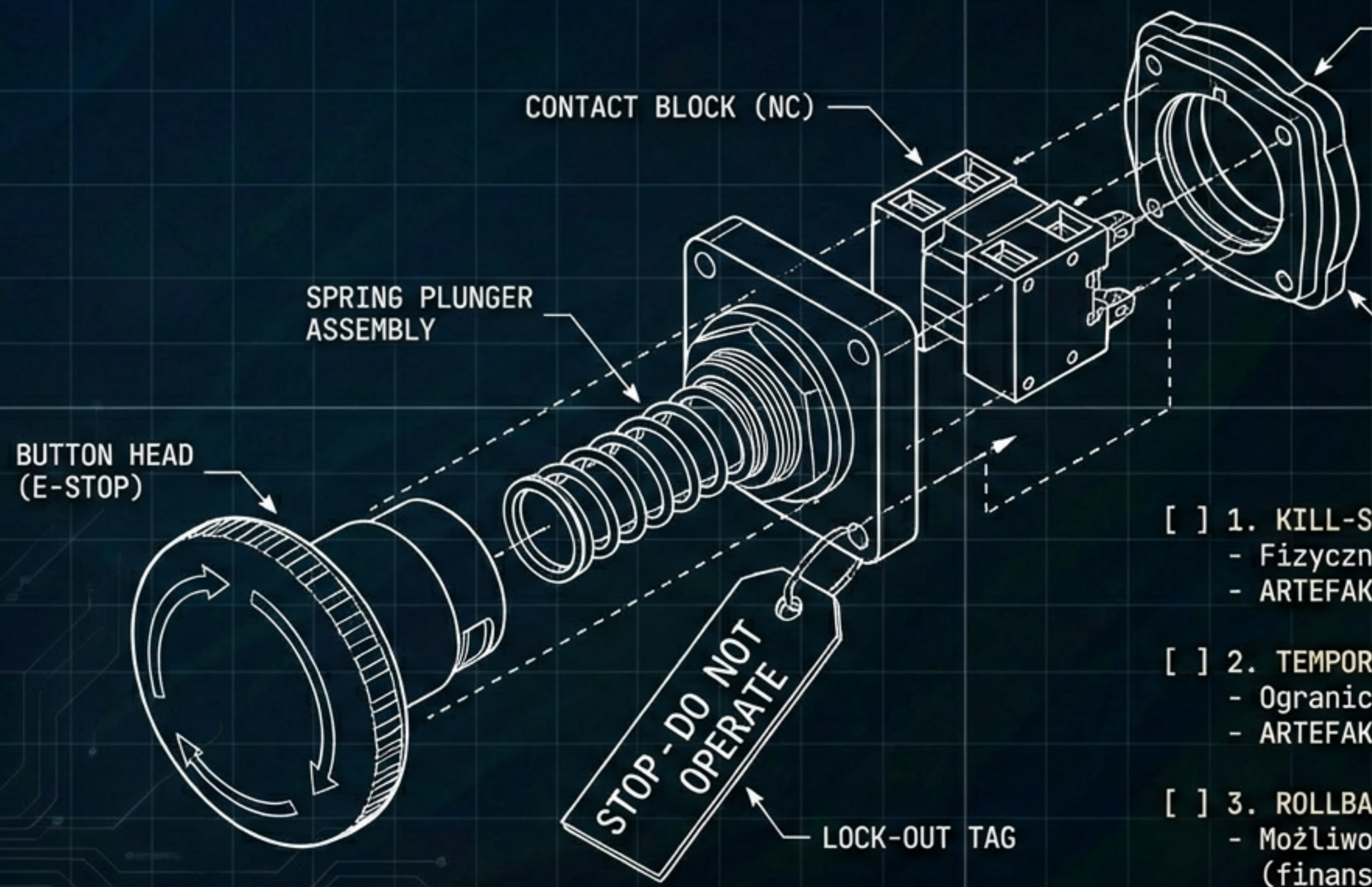
Filozofia "Minimum Architektonicznego"

OPTIMUM:
Maksymalna wydajność.
Minimalny koszt.
Ryzyko kaskady.

MINIMUM:
Warunek dopuszczalności.
Przestrzeń na bezpieczeństwo.
Hamulec bezpieczeństwa.

Minimum nie jest hamulcem postępu. Jest warunkiem dojrzałego postępu.

Mechanizmy Twarde: Hamulce i Bezpieczniki

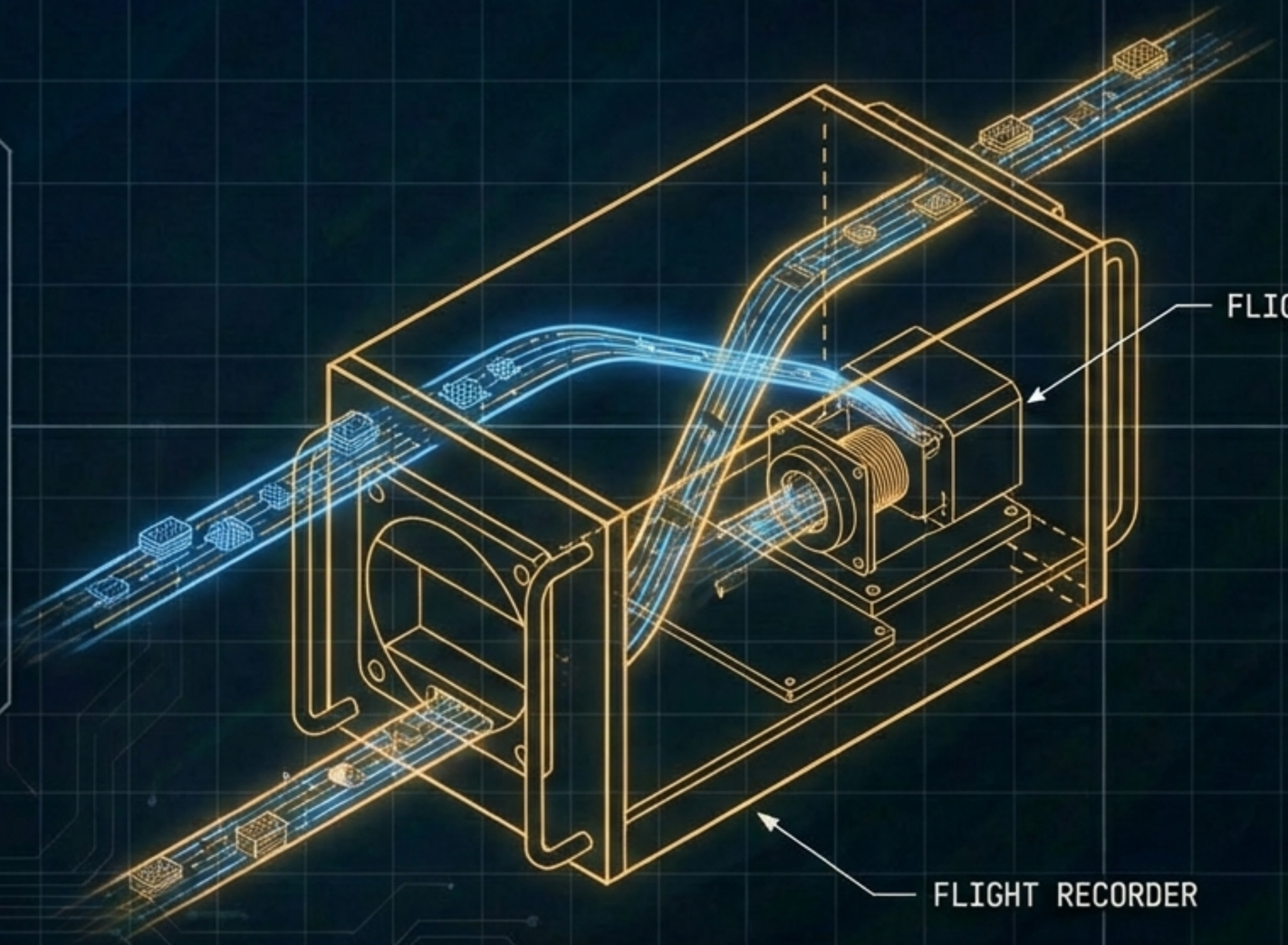


[ ] 1. KILL-SWITCH (Realny Wyłącznik)
- Fizyczny mechanizm zatrzymania.
- ARTEFAKT: Procedura STOP + Czas reakcji.

[ ] 2. TEMPORAL THROTTLING (Spowalnianie)
- Ograniczenie tempa do ludzkiej percepcji.
- ARTEFAKT: Specyfikacja buforów opóźnienia.

[ ] 3. ROLLBACK (Odwracalność)
- Możliwość cofnięcia skutków decyzji (finanse/prawo).
- ARTEFAKT: Logi cofania zmian.

Oczy i Uszy Systemu: Redundancja i Audyt

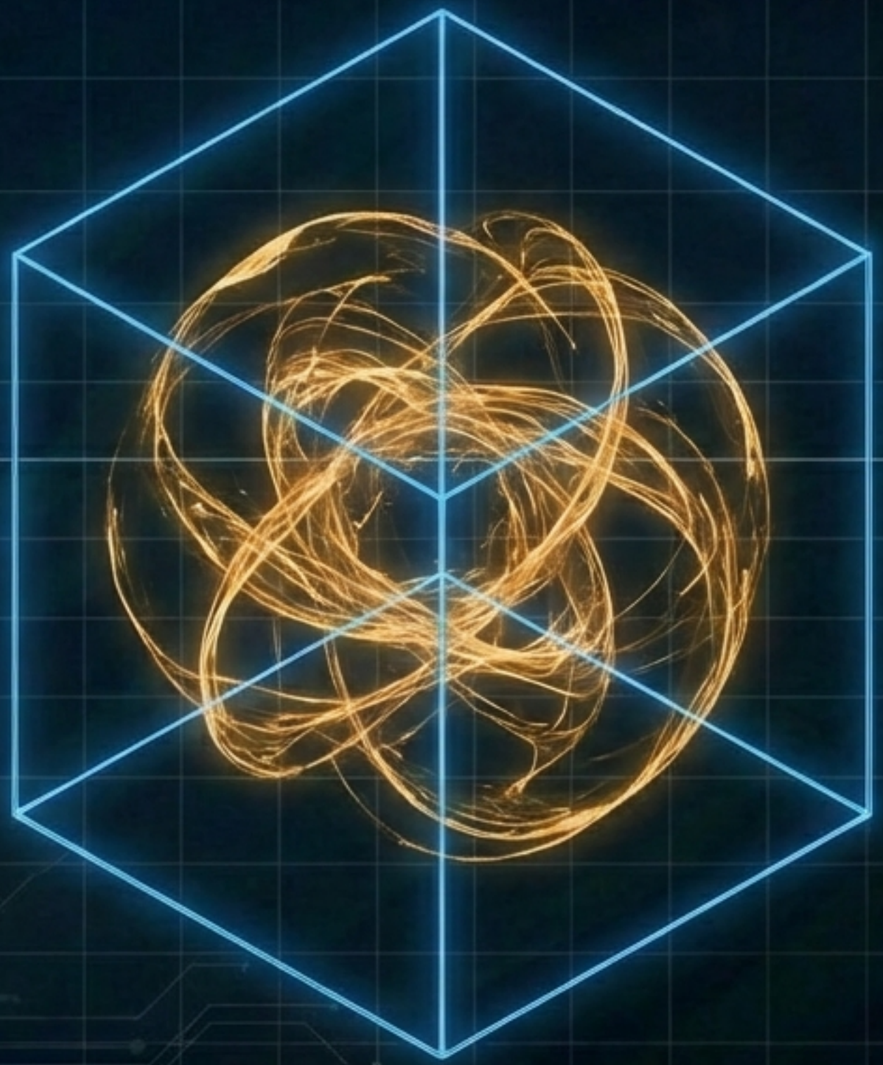


[ ] 1. REDUNDANCJA DECYZYJNA
- Decyzja wymaga potwierdzenia przez niezależny kanał.
- ARTEFAKT: Schemat kontroli krzyżowej.

[ ] 2. MONITORING ENTROPII I DRIFTU
- Wykrywanie nagłych zmian w rozkładzie decyzji.
- ARTEFAKT: Dashboard wskaźników stabilności.

[ ] 3. ŚLADY DECYZYJNE (Audytowalność)
- Kto, co, na jakiej podstawie?
- ARTEFAKT: Standard logowania i retencji.

Granice i Izolacja: Sandbox i Symulacja



[ ] 1. DIGITAL TWIN TESTING
- Stres-testy w symulacji przed wdrożeniem.
- ARTEFAKT: Raporty z symulacji 'czarnych łabędzi'.

[ ] 2. GRANICE KOMPETENCJI
- Sztywne ograniczenia tego, czego system NIE może robić.
- ARTEFAKT: Policy enforcement & Sandbox rules.

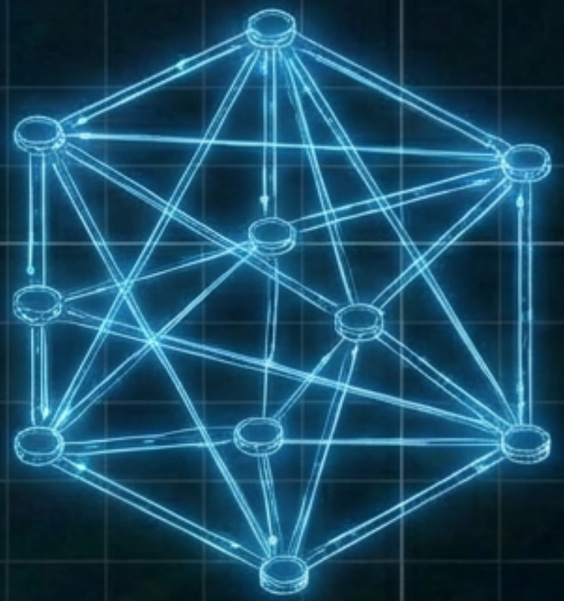
[ ] 3. MONITORING SKALI SPRZĘŻENIA
- Alarm, gdy system kontroluje krytyczny % rynku.
- ARTEFAKT: Dynamiczny wskaźnik udziału.

Niezależność: Przeciwdziałanie Lock-in

Ryzyko

Hub and Spoke

Cel



Mesh

[ ] 1. BRAK POJEDYNCZEGO PUNKTU AWARII
- Pluralizm dostawców (Multi-vendor).
- ARTEFAKT: Plan ciągłości działania (BCP).

[ ] 2. KNOWLEDGE PORTABILITY
- Eksport logiki i wag modelu.
- ARTEFAKT: Plan migracji technologicznej.

[ ] 3. ODPOWIEDZIALNOŚĆ (RACI)
- Konkretne nazwiska, nie 'zespoły'.
- ARTEFAKT: Macierz odpowiedzialności.

Gdzie stosujemy 'Minimum'? Definicja AI Wysokiego Wpływu

KRYTERIA:

[ ] 1. Infrastruktura Krytyczna (Energia, Transport)

[ ] 2. Prawa Obywatelskie (Zdrowie, Sprawiedliwość)

[ ] 2. Skala i Automatyzacja (Finanse, Administracja)

[ ] 3. Skala i Automatyzacja (Finanse, Administracja)

W tych obszarach standardy 'Move fast and break things' są niedopuszczalne.

Scenariusze Systemowe 2026–2035

Mapa Drogowa: Decyzje na 6–12 miesięcy

Prawo do Nieoptymalności

Optymalizacja nie może zastąpić podmiotowości.

- Technologia nie może zastąpić człowieka jako meta-systemu.
- Musimy zachować prawo do bycia mniej efektywnymi, aby pozostać wolnymi.
- Suwerenność > Efektywność.

Podsumowanie: Architektura Suwerenności

Granice kontroli nie są ograniczeniem postępu.
Są warunkiem przetrwania cywilizacji technicznej.

1. STABILNOŚĆ | 2. STEROWALNOŚĆ | 3. SUWERENNOŚĆ