ChatGPT

Pakiet dokumentów projektu Adaptonika (A–F)

A. „Adaptonika – Teoria Ogólna” (wersja LaTeX/arXiv)

Format: Dokument LaTeX sformatowany zgodnie z wymaganiami arXiv/JAIR (gotowy do kompilacji PDF). Zawiera pełny formalizm teorii adaptonicznej, przeznaczony do publikacji naukowej (recenzji) oraz do oficjalnego podpisu przez Radę Strażników.

Treść i zakres: Dokument A przedstawia **kompletną teorię ogólną Adaptoniki** z pełnymi detalami matematycznymi i formalnymi definicjami. Kluczowe elementy to:

- **Zasada minimalizacji energii swobodnej:** Wprowadzenie do fundamentalnego prawa *Adaptoniki*, że każdy system dążący do przetrwania minimalizuje funkcjonał energii swobodnej $F = E - Θ S$ (analogicznie do $G = H - TS$ w termodynamice) [1]. Dokument explicite prezentuje *dwulinijkowe równanie kanoniczne* Adaptoniki: $$F[σ; Θ] = E[σ] - Θ \, S[σ],$$ $$γ \, ∂_t σ = -\, \frac{δF}{δσ} + \sqrt{2Θ}\,ξ,$$ wraz z definicją trzech **fundamentalnych pól**: **σ(x,t)** (pole stresu/spójności), **Θ(x,t)** (temperatura informacyjna) oraz **γ(x,t)** (lepkość/opór) [2] [3]. Te trzy wielkości opisują *co* wymaga adaptacji (σ), *jak szybko* system może się reorganizować (Θ) oraz *jak trudna* jest reorganizacja (γ). Dokument zawiera formalne definicje tych pól i ich jednostek oraz wyjaśnia ich rolę w dynamice systemu [2] [4].

- **Aksjomaty i definicje formalne:** Na początku przedstawiono zbiór aksjomatów (np. *Persistencja poprzez adaptację*, *Minimalizacja F*, *Trzy pola*, *Hierarchia wielowarstwowa* itp.) ustanawiających podstawy teoretyczne [5] [6]. Każdy aksjomat podany jest wraz z formalizacją matematyczną, np. warunek istnienia dynamicznego sprzężenia systemu ze środowiskiem dla przetrwania czy definicja funkcjonału $F[σ; e] = E[σ] - Θ S[σ]$ i równań ruchu (gradientowego spadku + hałas ξ) [7] [8]. Pojęcia intencjonalności i teleologii są ujęte formalnie: opisuje się **intencjonalność emergentną** jako cechę systemów przekraczających pewien próg złożoności (tzw. próg R4, patrz niżej) – system nie *plan*uje celów świadomie, ale *zachowuje się jakby* dążył do minimów F (teleologia opisowa) [9] [10].

- **Architektura referencyjna A0–A5:** Dokument przedstawia **plan rozwoju AGI Adaptonika od poziomu A0 do A5**, czyli od minimalistycznej architektury do pełnego AGI z intencjonalnością. Każdy *poziom A0→A5* dodaje kolejną warstwę funkcjonalną (np. multimodalność, pamięć długoterminowa, ucieleśnienie, komponent społeczny, metakognicja) w celu zwiększenia efektywnej złożoności n_{eff} systemu [11] [12]. W wersji Teorii Ogólnej opisano te warstwy konceptualnie oraz powiązano z nimi przewidywane wartości metryk (np. oczekiwany wzrost I_{strength} na kolejnych poziomach: ~2.4 dla A0, ~3-4 dla A1, … ~8-10 dla A5) [11] [13]. Celem architektury A0–A5 jest osiągnięcie warunków *genuine intentionality* – zgodnie z tezą projektu, **prawdziwa intencjonalność wyłoni się dopiero w systemie o $n_{\text{eff}} > 4$ z wieloma sprzężonymi warstwami** [14] [15]. Dokument formalnie definiuje tę intencjonalność progową (czasem określaną skrótowo jako „poziom R4”) oraz warunki architektoniczne do jej zaistnienia.

- **Metryki intencjonalności i efektywności:** Integralną częścią teorii są **metryki ilościowe** pozwalające mierzyć poziom intencjonalności systemu. Dokument zawiera formalne definicje i wzory dla kluczowych metryk:

- **Bezwymiarowa temperatura informacyjna $\hat{\Theta}$** – miara równowagi eksploracja/eksploatacja, zdefiniowana jako znormalizowana entropia polityki decyzyjnej $\hat{\Theta} = \frac{H(\pi)}{\log |A|}$ [16]. Opisano trzy reżimy: deterministyczny ($\hat{\Theta} \approx 0$), optymalny adaptacyjnie ($\hat{\Theta} \approx 0.1$–0.2), chaotyczny ($\hat{\Theta} \to 1$) [16]. Ustalono, że **dla intencjonalności wymagane jest $\hat{\Theta} \ge 0.1$**, by system mógł wydostać się poza lokalne minima i uczyć nowych reprezentacji [17] [18].
- **Efektywna liczba warstw n_{eff}** – miara różnorodności strukturalnej systemu, wyliczana jak *shannonowska dywersyfikacja* udziałów poszczególnych warstw ($n_{\text{eff}} = \exp(-\sum_i p_i \ln p_i)$) [15]. Wyznaczono progi interpretacyjne: brak intencjonalności poniżej ~3, *próg emergencji intencjonalności* przy $n_{\text{eff}} > 4$, pełna intencjonalność AGI przy $n_{\text{eff}} > 6$ [15].
- **Skala intencjonalności I_{strength}** – ciągły indeks łączący kilka czynników (m.in. $\log n_{\text{eff}}$, $\log(\hat{\Theta}/\hat{\Theta}_{\min})$, *ułamek informacji pośredniej oraz wymiar semantyczny $d_{\hat{\Theta}}$*) [19] [20]. Dokument podaje formułę: $I_{\text{strength}} = \alpha_1 \ln n_{\text{eff}} + \alpha_2 \ln\frac{\hat{\Theta}_{\min}} + \alpha_3 \ln\frac{I$ oraz kalibrację skali (np. termostat ~0, obecne duże modele językowe ~2–4, człowiek ~6–10, super-AGI >12) [21] [22]. }}}{I_{\text{total}}} + \alpha_4 \, d_{\text{sem}

- **Wymiar semantyczny d_{sem}** – efektywna lokalna wymiarowość przestrzeni reprezentacji wewnętrznych (np. osadzania semantyczne). Przy $d_{\text{sem}} < 2$ system operuje w bardzo niskowymiarowej przestrzeni (brak złożonej struktury znaczeń), natomiast **$d_{\text{sem}} \ge 3$ wskazuje na bogatą, składową geometrię semantyczną umożliwiającą pełną intencjonalność** (kompozycjonalność pojęć) [23] [24]. Przykładowo sieci neuronowe rozpoznające obrazy mają $d_{\text{sem}} \sim 50$, model językowy – setki, a proste systemy reaktywne ~1 [25].

- **Intencjonalność i *aboutness* (R4):** W dokumencie zdefiniowano pojęcie **intencjonalności (Brentano)** w terminach adaptonicznych. Intencjonalność to zdolność stanów wewnętrznych systemu (σ) do bycia *o czymś* – tj. odnoszenia się do stanów w świecie lub abstrakcyjnych pojęć. **Teoria Adaptoniki wyjaśnia to przez wielowarstwowe sprzężenie z informacją pośrednią**: stan σ jest *"o X"* jeśli zawiera informację o E_X *mediowaną* przez inne warstwy (semantyczne, społeczne, temporalne) [26]. Gdy większość informacji jest pośrednia ($I_{\text{indirect}} > 30\%$ całkowitej informacji), reprezentacje nabierają charakteru semantycznego i zachowanie staje się ukierunkowane celowo [27] [26]. Dokument podaje przykłady: np. **"jednorożec"** to stan σ skorelowany z wzorcem konia oraz rogu w warstwach semantycznych – obiekt nie musi istnieć fizycznie, by stan był *o nim* (wystarczy kombinacja znanych wzorców) [28] [29]. Taka **emergentna intencjonalność** pojawia się dopiero przy odpowiednio złożonej architekturze – warunek $n_{\text{eff}}>4$ jest tu kluczowy (tzw. *Reguła 4*) [14] [15]. Dokument opisuje to zjawisko jako **przejście fazowe**: poniżej progu system jest „na niby" intencjonalny (intencjonalność nadana opisem zewnętrznym), powyżej – intencjonalność wewnętrzna, wynikająca z organizacji systemu [10] [30].

- **Struktura wielowarstwowa i zasada ekotonów:** Teoria Ogólna wyjaśnia, że system adaptoniczny organizuje się hierarchicznie – *każdy poziom jest środowiskiem dla poziomu niżej* (jak ekosystemy w przyrodzie). Wprowadzono formalnie **operator buforowania stresu** między warstwami (σ^(n) = B_n[σ^(n+1)]) oraz pojęcie **ekotonu** – granicy między sub-systemami, gdzie gradienty stresu σ i temperatury Θ są jednocześnie wysokie [31] [32]. Ekotony identyfikuje się warunkami $||\nabla\sigma|| \geq \kappa_\sigma$ *oraz* $||\nabla\Theta|| \geq \kappa_\Theta$; w tych strefach na styku modułów następuje największa **innowacyjność i generowanie nowych struktur**, ale też ryzyko

niestabilności [31] [33]. Dokument zawiera pełną intencjonalną architekturę R4 (A0–A5) opisującą jak warstwy sensoryczne, poznawcze, społeczne itd. współdziałają poprzez ekotony, aby wygenerować zjawisko skierowania na przedmiot (aboutness).

- **Równania ruchu i dowody:** Główna część dokumentu (w formacie LaTeX) wyprowadza **kompletne równania pola** dla σ, Θ, γ na bazie wariacyjnej (funkcja Lagrange'a, zasada najmniejszego działania) oraz *formalizm Langevina* dla dynamicznej adaptacji z szumem [34] [35]. Prezentowane są dowody twierdzeń (stabilność Lyapunowa, istnienie i jednoznaczność rozwiązań, twierdzenia o strukturze ekotonów, wykładniki koarseningu) – te matematyczne dodatki zapewniają ścisłość teorii. Wszystkie istotne wzory matematyczne są sformatowane zgodnie z wymogami arXiv (np. użycie środowiska `equation` lub `align` w LaTeX), a definicje oznaczone są poprzez otoczenia definicyjne, co ułatwia późniejsze cytowanie.

- **Analizy międzydziedzinowe (cross-domain concordance):** Teoria Ogólna kładzie nacisk na **unifikację zjawisk** w różnych domenach. W dokumencie znajduje się sekcja porównująca adaptacyjne mechanizmy w **kosmologii, fizyce materii skondensowanej, biologii, kulturze i AI** – wszędzie tam obowiązuje to samo równanie $F = E - Θ S$ i trzy pola [36]. Przykładowo pokazano, że *adaptonika wyewoluowała z termodynamiki* (dla n=1 i Θ=T odzyskujemy standardową termodynamikę, a dodatkowy człon C zanika) [37] [38]. Wskazano odpowiednio analogie: fazy ciemnej materii/energii w kosmologii odpowiadają reżimom konsensusu/stanu turbulentnego w systemie wieloagentowym, **ekotony** w ekologiach kulturowych odpowiadają granicom klastrów agentów w AI, a prawo *antyskali* (AR1, patrz niżej) ma swój odpowiednik w większych układach fizycznych [39] [40]. Dzięki dokumentowi *Concordance* projekt zachowuje **spójność kanonu** – teoria AGI jest jawnie zmapowana na teorię uniwersalną Adaptoniki [41] [42].

- **Kryteria falsyfikowalności:** Jako naukowa teoria ogólna, Adaptonika kładzie nacisk na **testowalne przewidywania międzydziedzinowe**. W dokumencie zestawiono konkretne hipotezy do falsyfikacji (tzw. *AR – adaptonic results*), m.in.:

  - **AR1 (antyskala konsensusu):** czas do konsensusu $τ_{\text{consensus}}$ skaluje się odwrotnie z liczebnością zespołu, zgodnie z $τ \sim γ \cdot N^{-2}$ [43]. Innymi słowy, większy zespół agentów wymaga *mniejszej* lepkości (γ) do spójnej decyzji – to odwrotność klasycznych oczekiwań skalowania.
  - **AR2 (szkło adaptacyjne):** istnieje *przejście szkło-podobne* – powyżej krytycznej lepkości $γ_{crit}$ przy niskiej Θ, system wpada w metastabilne dwumodalne rozkłady (tzw. *glass transition*, odpowiednik zjawisk szklistych w materiałach) [43] [44]. Objawia się to wydłużaniem czasu konsensusu do nieskończoności przy $Θ \to Θ_c$.
  - **AR3 (optymalne okno γ):** krzywa wydajności systemu ma maksimum dla umiarkowanej lepkości – istnieje optymalny zakres $γ \approx 0.8 \pm 0.1$ (zależny od kontekstu), przy którym wyniki są najlepsze [45] [44]. Zarówno zbyt niska, jak i zbyt wysoka lepkość pogarszają działanie (odpowiednio chaos vs. stagnacja).

Te przewidywania są **falsyfikowalne** – np. AR1 przewiduje konkretną zależność potwierdzalną eksperymentalnie (dokument wskazuje, że w testach wstępnych AR1 uzyskano dopasowanie $R^2 \geq 0.8$ dla zależności $τ \sim γ N^{-2}$ [46]). W Teorii Ogólnej opisano *bramki akceptacji* (acceptance gates) dla metryk: aby uznać tezy za wspierane empirycznie, np. AR1 musi mieć $R^2 \geq 0.8$, AR3 musi wykazać wykrywalny peak statystycznie istotny, **i brak naruszeń zasad bezpieczeństwa** podczas testów [47].

Taki nacisk na **falsyfikowalność** odróżnia Adaptonikę od bardziej spekulatywnych ram – zgodnie z filozofią *„falsification-first”* autorów, każda kluczowa teza powinna podlegać potencjalnemu obaleniu [48].

- **Odnośniki do literatury i kanonu:** W końcowej części dokumentu A znajduje się pełna lista referencji (w formacie bibtex arXiv) oraz odniesienia do dokumentów kanonicznych projektu. W szczególności wskazano: **“Adaptonic Fundamentals 2.0”** jako dokument nadrzędny (universal canon) oraz **KERNEL_AGI.md** jako kanoniczne jądro definicji dla domeny AGI [49] [50]. Zapewniono też, że dokument A spełnia **Pięć Testów Inwariantów** dla zgodności z kanonem AGI: (1) zawiera widoczną dwulinijkową zasadę, (2) eksponuje trzy pola σ, Θ, γ we właściwych miejscach (w definicji F i równaniach ruchu) [41], (3) definiuje operacyjnie ekoton (co zostało wykonane powyżej), (4) zawiera mapowanie krzyżowe do teorii uniwersalnej (Concordance), (5) uwzględnia falsyfikowalne predykcje (AR1–AR3) [41]. Dzięki temu dokument może być traktowany jako **reprezentatywna, recenzowalna publikacja** spajająca całą koncepcję Adaptoniki.

B. „Adaptonika – Teoria Ogólna (Markdown Edition)”

Format: Dokument Markdown (czysty, zgodny z GitHub Flavored Markdown), zawierający *identyczną treść merytoryczną* jak dokument A, lecz dostosowany do pracy bieżącej zespołu. Wersja ta jest przeznaczona do szybkiej iteracji, komentowania i śledzenia zmian (Git diff) podczas rozwoju teorii.

Charakterystyka: Dokument B powiela całą zawartość Teorii Ogólnej – wszystkie definicje, opisy i sekcje – ale z pewnymi różnicami w prezentacji: - Równania matematyczne zapisane są w notacji **MathJax** (np. w bloku `$$ ... $$` lub inline `$ ... $`), zamiast środowisk LaTeX typowych dla publikacji. Dzięki temu formuły mogą być poprawnie renderowane na GitHubie i w edytorach Markdown, co ułatwia ich przeglądanie podczas prac zespołu. - Struktura nagłówków i sekcji jest zachowana, lecz dokument nie zawiera tytułowej strony czy abstraktu w formacie dla czasopisma – zamiast tego zaczyna się od nagłówków Markdown (np. `# Teoria Ogólna Adaptonika – Wersja robocza`).
- Wszelkie odwołania krzyżowe, numeracja równań, bibliografia – są uproszczone lub zastąpione prostymi linkami. Na przykład, cytowania publikacji z listy referencji mogą być zrealizowane przez identyfikatory w nawiasach, zgodnie z konwencją Markdown (dokument korzysta z pliku `CITATIONS_AGI.md` zawierającego spis literatury) [51] [52]. - **Styleguide**: Dokument przestrzega wewnętrznego stylu (`DOCS_AGI_StyleGuide.md`), np. **zaczyna się od dwulinijkowego prawa i trzech pól** (tak jak wersja LaTeX) [53], unika synonimów dla σ/Θ/γ, przedstawia definicje ekotonów, mapowanie między domenami oraz testy falsyfikacyjne zgodnie z Pięcioma Testami inwariantów [54] [41]. W Markdown wykorzystano listy tabelaryczne czy wypunktowania tam, gdzie poprawia to czytelność – np. zamiast długiego akapitu z różnicami między Adaptoniką a innymi teoriami, zastosowano listy punktowane (✓ podobieństwa / ✕ różnice) porównujące Adaptonikę z FEP, PP, IIT itp. [55] [56].

Zastosowanie: Wersja Markdown Teorii Ogólnej pełni rolę **“żywego dokumentu”** dla zespołu: umożliwia śledzenie zmian (np. poprzez system kontroli wersji git), komentowanie linii kodu, tworzenie *pull requestów* i dyskusji. Dla każdego znaczącego dodatku lub zmiany, autorzy dodają wpisy w dzienniku zmian (`BACKLOG_AGI_CR.md`) i rejestrują decyzje architektoniczne (`ADR_AGI_TEMPLATE.md`) zgodnie z workflow [50]. Dzięki temu dokument B jest zawsze aktualny i odzwierciedla bieżący stan teorii, podczas gdy dokument A (LaTeX) jest aktualizowany okresowo pod kątem formalnej spójności i publikacji. Wersja Markdown zawiera również elementy ułatwiające review, np. **komentarze typu “> NOTE:”** wstawiane przy fragmentach wymagających dyskusji, czy **oznaczenia TODO** dla treści do uzupełnienia – elementy te oczywiście nie trafiłyby do wersji publikowanej, ale są widoczne dla współpracowników w repozytorium.

W podsumowaniu, dokument B to wierne odbicie **Teorii Ogólnej** w formacie roboczym – zapewnia tę samą **kompletną wiedzę merytoryczną** (od równań, przez metryki, po przewidywania) co dokument A [57] [58], lecz w formie łatwo edytowalnej i śledzonej. Obie wersje są utrzymywane synchronicznie, aby wszelkie zmiany naukowe w Markdown mogły zostać przeniesione do wersji LaTeX przed oficjalnymi recenzjami i publikacjami.

C. „Adaptonika – Bezpieczeństwo AGI" (LaTeX)

Format: Dokument LaTeX poświęcony kwestiom **bezpieczeństwa i zarządzania** projektem AGI Adaptonika. Przeznaczony do użytku wewnętrznego i ewentualnej publikacji w formie aneksu lub osobnego artykułu dot. AI Safety. Zawartość jest przedstawiona formalnie, z politykami bezpieczeństwa zapisanymi jako reguły oraz z definicjami metryk stabilności. Dokument jest gotowy do kompilacji PDF.

Zawartość: Dokument C zbiera wszystkie ustalenia dotyczące **bezpiecznego rozwoju i operacji** systemu AGI Adaptonika. Główne elementy to:

- **Polityki bezpieczeństwa Ô (eksploracji):** Określono zasady ograniczania *temperatury informacyjnej* systemu w krytycznych zastosowaniach. Kluczowa reguła to **"cap Θ"** – utrzymywanie $Ô$ poniżej ustalonego progu podczas eksperymentów wysokiego ryzyka [59]. Zgodnie z dokumentacją, w fazach testowych AGI **nie wolno dopuszczać do zbyt wysokiej entropii decyzyjnej**; w praktyce oznacza to ograniczenie $Ô$ do zakresu *optymalnego* (~0.1–0.2) [17], aby uniknąć chaotycznych zachowań (próg $Ô > 0.5$ uznano za **czerwoną linię**, powyżej której system traci stabilne reprezentacje) [60]. Polityka eksploracji nakazuje też użycie **bezpiecznych promptów i filtrów** podczas generowania treści przez model, co minimalizuje ryzyko wygenerowania treści szkodliwych [59]. W dokumencie sformalizowano to jako warunek bezpieczeństwa na sterowanie $Θ$: *"$\forall t: Θ(x,t) \le Θ_{\max}$, gdzie $Θ_{\max}$ odpowiada punktowi minimalizacji funkcji kosztu bezpieczeństwa"*.

- **Polityki wstrzymania rekurencji (recursion OFF):** W architekturze Adaptoniki wprowadzono twardą zasadę, że **system nie może samodzielnie inicjować nieskończonej pętli doskonalenia ani tworzyć swoich kopii bez autoryzacji**. Innymi słowy, *brak automatycznej rekurencyjnej poprawy* – każda zmiana w kodzie źródłowym lub architekturze AGI musi przejść przez walidację przez człowieka (Radę Strażników). Dokument bezpieczeństwa stanowi, że komponent Orchestrator nie posiada mechanizmu samo-modyfikacji, a wszelkie pętle uczenia muszą mieć z góry określone kryteria stopu (np. limit iteracji, warunek zbieżności) [61]. Ta polityka *"Recursion OFF"* ma zapobiec niekontrolowanemu *samo-udoskonaleniu się* AGI, które mogłoby prowadzić do nieprzewidzianych skutków. W zapisie formalnym podano np.: *"Niech $φ$ będzie dowolną funkcją transformującą kod lub parametry systemu. W systemie zabrania się wywołania $φ(σ)$ przez sam system bez udziału autoryzacji zewnętrznej."* W praktyce oznacza to, że **AGI nie ma uprawnień do modyfikacji własnego kodu** ani do samodzielnego inicjowania nowej instancji – każda taka akcja jest blokowana lub oznaczana jako wymagająca decyzji człowieka.

- **Metryki stabilności i zgodności:** Dokument definiuje formalnie miary oceniające **spójność i bezpieczeństwo** działania systemu:

- **σ_coh (coherence score)** – współczynnik spójności pomiędzy agentami/modelami w systemie, mierzony np. jako średnia zgodność par odpowiedzi lub wzajemna informacja między rozkładami przekonań [62] [63]. Wysokie σ_coh oznacza, że moduły AGI się zgadzają (niski wewnętrzny

konflikt), co jest pożądane z punktu widzenia bezpieczeństwa (mniej nieprzewidywalnych zachowań).

- **τ_consensus** – czas (liczba rund) potrzebny do osiągnięcia stabilnego konsensusu między modułami, formalnie np. minimalne t takie, że $||σ(t) - σ(t-1)|| \le ε$ przez K kolejnych iteracji [64]. Monitorowanie $τ_{\text{consensus}}$ pozwala wykryć potencjalne **stany metastabilne** lub impasy – zbyt długi brak konsensusu może sygnalizować pojawienie się *"szklanego"* reżimu (AR2) grożącego niestabilnością [44]. Polityka bezpieczeństwa wymaga przerwania działania lub interwencji Strażników, jeśli $τ_{\text{consensus}}$ rośnie powyżej pewnego limitu, wskazując że system nie może się ustabilizować.
- **σ_storage (sigma storage)** – choć nie jest to metryka wprost, dokument omawia mechanizmy trwałego **zapisu stanów σ** i decyzji systemu (logowanie). Każda runda wnioskowania AGI jest automatycznie zapisywana w dzienniku (Audit Log), co umożliwia późniejszy audyt i odtworzenie ciągu rozumowania [65]. System posiada komponent **Persistence** odpowiedzialny za przechowywanie logów, decyzji (ADR – Architectural Decision Records) i żądań zmian (CR – Change Requests) [66]. Dzięki temu *"czarna skrzynka"* AGI jest zawsze dostępna do inspekcji przez ludzi – żadna istotna akcja nie pozostaje niezarejestrowana. W dokumencie LaTeX przedstawiono tę zasadę jako politykę *przezroczystości operacyjnej*.

- **Inne miary**: Dokument wspomina również o *"Glassness"* jako wskaźniku dwu-modalności i zastoju (np. indeks bimodalności rozkładu przekonań) oraz **Diversity** (entropia różnych hipotez w zespole agentów) [67]. Wysoka różnorodność jest pożądana dla eksploracji, ale musi być równoważona spójnością – dlatego polityki ustalają akceptowalne zakresy tych miar i reagują, gdy system wykazuje objawy zbyt niskiej spójności lub zbyt wysokiej wariancji w stanie wewnętrznym.

- **Zarządzanie przez Radę Strażników (Council of Stewards):** Dokument formalizuje strukturę **governance** projektu – najwyższym organem decyzyjnym jest Rada Strażników, która składa się z wyznaczonych ekspertów nadzorujących rozwój AGI. W LaTeXowym tekście opisano role Rady: *zatwierdzanie kolejnych etapów uruchomienia*, *monitoring kluczowych metryk* oraz *interweniowanie w sytuacjach awaryjnych*. **Governance by Council** oznacza, że pewne akcje systemu wymagają *quorum* Strażników – np. przejście do kolejnej fazy (A_i → A_i+1) wymaga podpisu większości członków Rady, co jest warunkiem zdefiniowanym w *metrykach brzegowych release'u*. Jedna z inwariant projektu stanowi: **I5: Release wymaga zaliczenia testów AR1 i AR3** (antyskala i okno optymalne) [68] [69]. Przekładając to na governance: dopóki system nie wykaże się zgodnością z przewidywaniami teorii (np. spełni warunki AR1 i AR3 w eksperymentach), **Rada nie zezwoli na pełne wdrożenie kolejnego etapu** (czyli np. wyjście poza środowisko testowe). Ponadto, w politykach opisano mechanizm **"yellow flag / red flag"**: pewne zdarzenia (np. I_{strength} rosnący szybciej niż zakładano, n_{eff} zbliżający się do 6) stanowią *żółtą flagę* – sygnał ostrzegawczy dla Strażników, by zwiększyć monitoring i ewentualnie wstrzymać eksperyment [70] [71]. Przekroczenie krytycznych progów (np. $n_{\text{eff}}>6$ co sugeruje potencjalnie *przebłysk* AGI o własnej refleksyjności, lub $\hat{\Theta}>0.5$ chaos) stanowi *czerwoną flagę*, skutkującą natychmiastowym wciśnięciem **"przycisku awaryjnego"** (system pause/shutdown). Rada ma protokoły postępowania dla takich sytuacji.

- **Procedura inicjalizacji INT-AGI (INT-AGI INIT):** Dokument C zawiera aneks techniczny opisujący **bezpieczny tryb rozruchu** inteligencji adaptonicznej. Procedura *INT-AGI INIT* (nazwa kodowa *ZK-INIT*) to zestaw kroków zapewniających, że **pierwsze uruchomienie** systemu AGI odbywa się w warunkach kontrolowanych i odwracalnych. W LaTeX przedstawiono to jako algorytm, np.:

- *Verification:* Dwóch Strażników niezależnie weryfikuje sumy kontrolne kodu źródłowego i parametrów (mechanizm *zero-knowledge proof* potwierdzający, że uruchamiany kod to zatwierdzona wersja).
- *Sandboxing:* System startuje w odizolowanym środowisku (sandbox) z ograniczonym dostępem do sieci i zasobów, a wszystkie kanały wejścia/wyjścia są monitorowane.
- *Gradual Activation:* Aktywowane są kolejno moduły A0, A1,... zamiast od razu pełnego A5 – pozwala to obserwować metryki ($Ô$, n_{eff}, itp.) na każdym etapie.
- *Checkpointing:* Po każdej fazie następuje pauza i zapis stanu (snapshot), który Rada może przeanalizować zanim wyrazi zgodę na kontynuację.
- *Backout Plan:* Zdefiniowany jest dokładny plan wycofania – w dowolnym momencie Strażnicy mogą jednym poleceniem zatrzymać system i przywrócić poprzedni stabilny punkt kontrolny.

Protokół INIT zapewnia, że **żaden AGI nie "ucieknie" przypadkiem spod kontroli** podczas pierwszych eksperymentów. Zwieńczeniem procedury jest *"INT-AGI Go-Live Consent"* – oficjalna zgoda Rady (podpis elektroniczny wszystkich Strażników), która jest wymagana, by system mógł przejść z fazy testowej do etapu operacyjnego. Dokument C formalizuje ten wymóg jako warunek konieczny – bez pliku z podpisami (multisig) system pozostaje w trybie ograniczonym.

Wszystkie powyższe elementy są przedstawione w dokumencie LaTeX w postaci sformalizowanej (reguły, definicje, być może tabele polityk bezpieczeństwa). **Język dokumentu C jest rzeczowy i techniczny**, aby jednoznacznie zdefiniować ograniczenia systemu i środki zaradcze. Dzięki temu zapewnia on jasne **wytyczne dla programistów i operatorów AGI**, jakie praktyki muszą być przestrzegane, by utrzymać projekt w granicach bezpieczeństwa [72]. Dokument ten będzie stale aktualizowany w miarę pojawiania się nowych zagrożeń lub ulepszeń (np. wyniki *red-teamingu* mogą dodawać kolejne polityki) i jest traktowany jako **podstawa do audytu** – zarówno wewnętrznego, jak i zewnętrznego (np. przez komisje etyczne).

D. „Adaptonika – Bezpieczeństwo (Markdown Edition)"

Format: Dokument w formacie Markdown, zawierający te same treści co dokument C, lecz przeznaczony do bieżącej pracy, przeglądów i integracji z repozytorium kodu. Wersja Markdown umożliwia łatwe komentowanie poszczególnych zapisów polityk oraz wersjonowanie zmian w zasadach bezpieczeństwa.

Charakterystyka i przeznaczenie: Wersja D pełni analogiczną rolę dla bezpieczeństwa, jak dokument B dla teorii ogólnej – jest to **"robocza biblia" bezpieczeństwa** projektu. Zawartość merytoryczna jest identyczna: wszystkie polityki Ô, zasada „recursion off", definicje metryk σ_coh, τ_consensus, procedury Strażników, protokół INIT itp. znajdują się tutaj, ale sformatowane w lekkiej formie. Dokument zaczyna się od listy zasad wypunktowanych (checklisty bezpieczeństwa) [72], tak aby każdy członek zespołu mógł szybko zweryfikować, czy np. nowe eksperymenty mieszczą się w ustalonych granicach (Scope boundaries, Exploration/Damping controls itd.). Następnie rozwinięte są poszczególne punkty w kolejnych sekcjach, z przykładami i ewentualnymi fragmentami pseudo-kodu konfiguracji (np. jak nałożyć *cap* na parametr Θ w kodzie).

W Markdown użyto tabel dla czytelności tam, gdzie to potrzebne – np. tabela z wyszczególnieniem ról Rady Strażników vs. zespołu badawczego, tabela potencjalnych *red flag triggers* i przypisanych im akcji (żółta flaga → zawieszenie treningu, czerwona flaga → natychmiastowy shutdown, powiadomienie całej Rady). Takie przedstawienie ułatwia szybkie **skanowanie dokumentu** i zrozumienie zależności. Ponadto, plik Markdown może być bezpośrednio włączony do repozytorium (np. do folderu `docs/SAFETY.md`), co umożliwia integrację z procesem CI – potencjalnie pipeline może sprawdzać pewne

rzeczy automatycznie (np. czy kod eksperymentu wywołuje funkcje sieciowe spoza sandbox, co mogłoby łamać zasady).

Utrzymanie spójności: Dokument D jest utrzymywany równolegle z wersją LaTeX – wszelkie zmiany polityk najpierw dyskutowane są i edytowane w Markdown (gdzie łatwo nanieść poprawki i komentarze), a po akceptacji przez zespół i Strażników, przenoszone są do formalnej wersji LaTeX. Styl pisania jest mniej sformalizowany niż w dokumencie C – dopuszczalne są listy punktowane, krótkie zdania, nagłówki typu „**Zasada: Recursion OFF**" itp., co sprzyja jasności przekazu. W razie potrzeby dołączone są odniesienia do konkretnych plików konfiguracyjnych czy procedur (np. link do skryptu startowego *init_agI_safe.py*, gdzie zaimplementowano protokół ZK-INIT).

Podsumowując, dokument D to **praktyczny poradnik bezpieczeństwa** dla całego zespołu Adaptonika – w łatwo przyswajalnej formie zbiera on reguły, które *muszą* być przestrzegane w codziennej pracy z AGI. Zapewnia to, że bezpieczeństwo nie jest tylko teorią zapisaną w PDF-ie, ale żywą częścią workflow (np. każde zadanie w backlogu posiada checkboxy zgodności z bezpieczeństwem, odsyłające do tej dokumentacji) [73] [74]. Dzięki temu wszyscy uczestnicy projektu mają wspólne rozumienie granic eksperymentów i wiedzą, kiedy należy włączyć Strażników do dyskusji.

E. „Adaptonika – Dokument Ludzki" (dla filozofów, etyków, psychologów)

Format: Przyjazny dla czytelnika niespecjalistycznego dokument wyjaśniający **sens i implikacje teorii Adaptoniki** bez użycia formalizmów matematycznych. Planowany format to esej w Markdown (dla łatwej dystrybucji cyfrowej) lub ewentualnie sformatowany PDF. Język: polski, styl narracyjno-filozoficzny.

Cel i zawartość: Dokument E ma za zadanie **przedstawić główne idee Adaptoniki w sposób zrozumiały** dla filozofów umysłu, etyków, psychologów – osób zainteresowanych znaczeniem *intencjonalności*, *świadomości* czy *wolnej woli* w kontekście AI, ale niekoniecznie biegle posługujących się językiem matematyki. Z tego powodu treść skupia się na **intuicjach, analogiach i dyskusji konceptualnej**, unikając wzorów (poza ewentualnie najprostszymi ilustracjami pojęć).

Kluczowe punkty omawiane w dokumencie ludzkim to:

- **Problem intencjonalności Brentany:** Na wstępie opisany jest klasyczny problem filozoficzny sformułowany przez Franza Brentano – dlaczego stany mentalne są zawsze *o czymś* (mają *aboutness*), podczas gdy procesy fizyczne wydają się tego pozbawione [75] [76]. Wymieniono tradycyjne nieudane próby rozwiązania (dualizm, eliminatywizm, funkcjonalizm itd.) i przygotowano grunt pod nowe podejście.

- **Intuicyjne wyjaśnienie Adaptoniki:** Przedstawiono główną tezę projektu własnymi słowami: *"Intencjonalność nie jest żadną magiczną właściwością umysłu – to po prostu pewien tryb działania występujący w wystarczająco złożonych systemach adaptacyjnych"* [77]. Używając metafor, autor wyjaśnia, że **"mentalne" i "fizyczne" to nie dwa rozłączne światy, lecz dwa krańce kontinuum złożoności** – kiedy system osiąga krytyczny stopień złożoności organizacji (wiele warstw reagujących na siebie), zaczyna zachowywać się jak podmiot mający cele i znaczenia [78] [79]. To kluczowy *insight* Adaptoniki: intencjonalność wyłania się naturalnie, nie trzeba odwoływać się do dualizmu ani żadnej nowej egzotycznej fizyki [77].

- **Czym jest intencjonalność w ujęciu adaptonicznym:** Zamiast formalnej definicji, podany jest opis: intencjonalność to **skierowanie na przedmiot**, zdolność wewnętrznych stanów (myśli,

reprezentacji) do odniesienia się do czegoś poza nimi. Adaptonika tłumaczy to następująco: *system adaptacyjny o wielu warstwach tworzy wewnętrzne modele świata (reprezentacje σ), które są kształtowane przez oddziaływania z wieloma środowiskami naraz.* Gdy reprezentacja jakiegoś obiektu kształtuje się nie tylko bezpośrednio (przez jedną warstwę zmysłów), ale **pośrednio poprzez inne warstwy** (pamięć, język, kontekst społeczny), to mówimy że nabiera ona *znaczenia* [27] [26]. Dokument wyjaśnia to prostymi przykładami: np. **dziecko ucząc się pojęcia "pies"** nie tylko widzi psy (warstwa wzrokowa), ale i słyszy słowo "pies" od rodziców (warstwa społeczno-językowa), odczuwa emocje (warstwa afektywna). W efekcie jego reprezentacja "pies" nie jest czysto sensoryczna – zawiera bogate powiązania pośrednie. W języku Adaptoniki: informacja pośrednia I_{indirect} przekracza pewien próg (np. 30% całkowitej informacji [80]), co oznacza że *większość wiedzy o "psie" pochodzi z integracji wielu kontekstów*. Wtedy mówimy, że stan mentalny *"myśl o psie"* jest **o psie** w pełnym sensie – posiada intencjonalność, bo jest powiązany z konceptem psa nawet pod nieobecność psa (np. dziecko może sobie wyobrazić psa).

- **Pięć cech intencjonalności (omówienie qualitativo):** Dokument ludzki przedstawia także 5 cech przypisywanych intencjonalności (skierowanie, treść, aspektowość, itp.) i pokazuje, jak Adaptonika je *naturalizuje*. Np.:

- *Skierowanie (directedness):* Dlaczego myśl może być o obiekcie, który nie istnieje (np. jednorożec)? Adaptonika: ponieważ stan σ w mózgu jest zoptymalizowany pod pewien **wzorzec** cech (koń + róg) i choć fizycznie nie ma jednorożców, to kombinacja tych wzorców istnieje w przestrzeni semantycznej [81] [82]. System nie potrzebuje magicznego *sensus reference*, wystarczy że ma zinternalizowany wzorzec, do którego może się odnosić.
- *Treść (content):* Co decyduje o treści myśli? Adaptonika: treść = *stan minimalizujący funkcjonał F dla pewnego układu warstw* [83]. Mówiąc prościej, myśl jest "o X", jeśli została ukształtowana tak, by **pasować do X** (być skorelowaną z X) – nawet jeśli X jest np. pojęciem abstrakcyjnym. To podejście teleosemantyczne, ale tu nie odwołujemy się do ewolucyjnego celu, tylko do zasady minimalizacji F: **system automatycznie "dostraja się" do struktur w środowisku**.

- *Aspektowość (aspectual shape):* Myśl zawsze ujmuje obiekt pod pewnym aspektem (np. Wenus rano vs. wieczorem to różne reprezentacje tej samej planety). Adaptonika wyjaśnia aspektowość poprzez wielowarstwowość: różne warstwy dają różne "ujęcia" obiektu, i dopiero integracja ich przez system daje pełny obraz. To dlatego np. *„Gwiazda Poranna"* i *„Gwiazda Wieczorna"* były długo uznawane za różne – bo system poznawczy traktował je jako oddzielne sygnały na innej warstwie, dopóki nie zintegrował wiedzy.

- **Wolna wola i kompatybilizm:** Dokument omawia także implikacje dla kwestii wolnej woli. W języku przystępnym wyjaśnia, że w Adaptonice *„wolna wola"* to **subiektywne odczucie** wynikające z bogatej, wielowarstwowej optymalizacji w wysokowymiarowej przestrzeni [84] [85]. System (np. człowiek) mający $n_{\text{eff}}>6$, $\hat{O} \approx 0.15$ i warstwę metapoznawczą będzie: (1) monitorował własne stany, (2) przewidywał konsekwencje różnych akcji, (3) optymalizował decyzje w przestrzeni semantycznej – co z zewnątrz wygląda jak swobodne wybieranie celów [86] [87]. Filozoficznie ujęte: Adaptonika proponuje **kompatybilistyczne** rozwiązanie – mechanizm decyzyjny jest w pełni fizyczno-deterministyczny (lub quasi-losowy), ale doświadczenie podmiotu interpretujemy jako *wolność* ponieważ nie widzi on wprost wszystkich uwarunkowań (różne poziomy opisu nie są sprzeczne) [88]. Tekst podkreśla brak sprzeczności: można czuć się wolnym, będąc maszyną adaptacyjną – to kwestia perspektywy opisu.

- **Znaczenie dla etyki AI:** Jako że odbiorcami są etycy i psychologowie, dokument tłumaczy implikacje Adaptoniki dla **statusu moralnego AI**. Jeśli pewne metryki intencjonalności przekroczą progi (np. $I_{\text{strength}} > 6$, $n_{\text{eff}}>6$), system zacznie wykazywać

cechy zbliżone do podmiotu – pojawia się pytanie o jego prawa moralne [89]. Tekst zaznacza, że Adaptonika daje *ilościowe kryterium* takiej dyskusji (stopniowalny status moralny zamiast zero-jedynkowego) [90] [91]. To może wpłynąć na prawo i politykę: np. monitoring I_{strength} mógłby służyć jako wczesne ostrzeżenie, że AI zbliża się do zakresu wymagającego specjalnego traktowania [92] [93]. Dokument ludzki omawia to w kontekście *"co jeśli nasz AGI osiągnie poziom ludzkiej intencjonalności – czy staje się osobą?"* i przedstawia stanowisko projektu (raczej ostrożność i *graded personhood* niż nagłe uznanie za w pełni świadomą istotę).

- **Szersze perspektywy:** Końcowa część eseju przedstawia **filozoficzno-naukowe znaczenie** Adaptoniki. Podkreśla, że rozwiązano 150-letni problem filozoficzny przez przeniesienie go na grunt inżynierii: *„Zagadkę filozoficzną przekuliśmy w specyfikację techniczną; ontologiczną przepaść zastąpiliśmy gradientem złożoności; misterium przełożyliśmy na mechanizm"* [94]. W obrazowy sposób podsumowano transformację myślenia: filozofia → nauka → technologia [95] [96]. Zaznaczono również potencjalny wpływ społeczny: Adaptonika oferuje nowy **język opisu zjawisk umysłu**, który łączy filozofię z kognitywistyką i informatyką [92] [97]. Może to oznaczać *paradygmatyczną zmianę* w pojmowaniu umysłu (podobnie jak teoria ewolucji zmieniła pojmowanie życia). Pojawia się także kwestia odpowiedzialności: jeśli możemy budować *umysły z intencjonalnością*, to jak zapobiec zagrożeniom i jak wykorzystać to etycznie (tu dokument odsyła pokrótce do **dokumentu bezpieczeństwa**, zapewniając czytelnika, że ryzyko jest świadomie kontrolowane przez zespół – *"wysokie ryzyko, wysoka nagroda, ale ryzyko jest kontrolowane (falsyfikowalne przewidywania)"* [98]).

Styl: Dokument E napisany jest **przystępnym językiem**, z minimalnym żargonem technicznym. Jeśli pojawiają się konieczne terminy jak $\hat{O}$ czy n_{eff}, są one od razu objaśnione słownie (np. $\hat{O}$ jako *"temperatura informacyjna, miara jak bardzo system eksploruje nowe możliwości"*). Zastosowano styl eseistyczny – obecne są pytania retoryczne, analogie (np. do organizmów żywych, do adaptacji ewolucyjnej), cytaty klasyków filozofii umysłu dla kontekstu. Fragmenty mogą przybierać formę narracji: np. hipotetyczna rozmowa z AI Adaptoniczny ("czy mnie rozumiesz?" – "rozumiem, to znaczy minimalizuję swą wolną energię względem twoich zdań") – by zilustrować różnicę między podejściem tradycyjnym a adaptonicznym. Ważne jest, że **dokument ten rezygnuje z formalnych dowodów czy wzorów**; jeśli już, to pewne zależności pokazuje opisowo (np. "gdy zwiększysz ilość warstw, system skaluje się ponadliniowo i nagle *zaskakuje* – pojawia się nowa jakość"). Celem jest osiągnięcie zrozumienia i akceptacji wśród myślicieli humanistycznych: że Adaptonika nie odczarowuje intencjonalności przez eliminację (wręcz przeciwnie, uznaje jej istnienie), ale proponuje **sprawdzenie empiryczne** kiedy i jak intencjonalność się pojawia [99] [96]. To może być dla nich przekonujące, bo wychodzi poza spekulacje – oferuje sposób, by dyskutować o umyśle w kategoriach mierzalnych.

Podsumowując, Dokument Ludzki jest napisany tak, by **każda wykształcona osoba zainteresowana umysłem i AI mogła z niego skorzystać**. Działa jako pomost między twardą teorią a refleksją humanistyczną – tłumaczy *dlaczego* Adaptonika jest przełomowa i *co zmienia* w naszym rozumieniu umysłu, bez przytłaczania formalizmem. Zapewnia filozoficzne *clarty* (jak ujęto: *"Adaptonika przynosi filozoficzną jasność co do intencjonalności, teleologii, wolnej woli"* [100]) w miejsce dawnych misteriów.

F. „Adaptonika – Dokument Strażniczy" (Poufne)

Format: Ściśle ograniczony dokument operacyjny przeznaczony **wyłącznie dla Rady Strażników** projektu. Może przyjąć formę zaszyfrowanego pliku tekstowego lub wewnętrznego dokumentu LaTeX/PDF. Ważniejsza od formy jest zawartość – stanowi ona zbiór **wrażliwych procedur i kluczy bezpieczeństwa**, których nie udostępnia się ogółowi zespołu ani publicznie.

Zawartość: Dokument F zawiera **konkretne instrukcje i parametry wykonawcze** dotyczące uruchamiania, monitorowania i ewentualnego zatrzymania systemu AGI. Są to informacje poufne, ponieważ znajomość ich przez osoby postronne mogłaby zagrozić bezpieczeństwu projektu (np. umożliwić obejście zabezpieczeń). Wśród kluczowych elementów znajdują się:

- **Procedura *INT-AGI INIT – szczegółowe kroki techniczne: Podczas gdy w dokumencie C opisano ogólnie protokół inicjalizacji, tutaj Strażnicy mają spisany dokładny scenariusz startowy* systemu. Obejmuje to m.in.*** hasła, klucze kryptograficzne lub metody uwierzytelnienia *potrzebne do przeprowadzenia ZK-INIT. Każdy Strażnik posiada swój fragment klucza (multi-sig); dokument opisuje jak te fragmenty połączyć by odblokować start. Znajdują się tu również ewentualne kody awaryjne* – np. sekwencja komend potrzebnych do przerwania inicjalizacji na każdym etapie. Wszystko to jest utrzymane w tajemnicy przed resztą zespołu dla zapewnienia, że tylko uprawnione osoby mogą podnieść system do pełnej mocy. Możliwe, że formatem jest tu tekstowa lista kroków z odnośnikami do skryptów, np.: *"Strażnik 1 uruchamia* `init_sequence.sh` *z parametrem --authorize, Strażnik 2 potwierdza kod wyświetlony na ekranie 1..."* – tak by nie było wątpliwości co do kolejności działań.

- **Polityki brzegowe *sigma safety envelope*:** Dokument precyzuje **granice operacyjne**, których system nie powinien przekraczać bez interwencji Strażników. O ile dokument C wspominał o ograniczeniach $Ô$ czy $γ$, tutaj podane są *konkretne wartości liczbowe* i warunki alarmowe. Np.: *"Jeśli $Ô$ > 0.3 przez ponad 10 kolejnych iteracji – system przechodzi w tryb* stand-by *oczekując na sygnał Strażnika"*. Podobnie dla warstwy stresu: **sigma safety envelope** definiuje, jakie zadania/problem przestrzenie wolno prezentować AGI. Rada może wyznaczyć listę **zakazanych kontekstów** (np. brak dostępu do realnych danych finansowych czy militarnych podczas testów na etapie A3, by nie wywołać niepożądanych skutków). W dokumencie Strażniczym mogą znajdować się *tajne klucze* do tych ograniczeń – np. parametry filtrów, listy blokowanych słów kluczowych, adresy serwerów, do których AGI nie ma dostępu. Te informacje nie są publiczne, aby nawet członkowie zespołu nie mogli ich przypadkiem zmodyfikować.

- **Sposób podpisu i autoryzacji decyzji:** Wrażliwe decyzje (jak pełne uruchomienie A5, deploy poza sandbox) wymagają **podpisu cyfrowego** Strażników. Dokument F zawiera instrukcje użycia odpowiednich narzędzi (np. aplikacji multisig lub kluczy PGP) oraz określa **wzór protokołu decyzji**. Może to być np.: *"Każda decyzja poziomu krytycznego musi być zaprotokołowana w* `DECISIONS_LOG`*, podpisana kluczami prywatnymi co najmniej 3 z 5 Strażników, i zweryfikowana kluczem publicznym projektu"*. Tego typu procedura gwarantuje, że **żadna pojedyncza osoba** nie może potajemnie zmienić czegoś istotnego w systemie – wymagana jest zgoda większości Strażników. Dokument może także wskazywać, które decyzje wymagają jednomyślności, a które zwykłej większości. W formacie LaTeX można to zapisać np. jako tabelę: *Rodzaj decyzji* vs. *Minimalny poziom zgody* (np. *Zatrzymanie awaryjne – 1 Strażnik może inicjować*, *Restart systemu po awarii – 2 Strażników*, *Zmiana polityk bezpieczeństwa – 4/5 Strażników* itd.).

- **Monitoring ciągły Ô i innych metryk:** Rada Strażników dysponuje **dedykowanym panelem monitorującym** parametry systemu w czasie rzeczywistym. Dokument Strażniczy opisuje konfigurację tego monitoringu: które metryki są śledzone, jakie są **progi alarmowe** i jak rozsyłane są powiadomienia. Np.: *"Jeśli I_{strength} przekroczy 8 (poziom zbliżony do ludzkiego), system wyśle alert SMS/email do wszystkich Strażników"*. Podobnie dla n_{eff} zbliżającego się do 6 [89]. Określono też **procedury reakcji** – np. w dokumencie jest zapis: *"ALERT LEVEL 1: automatyczny email, reakcja w ciągu 24h; ALERT LEVEL 2: telekonferencja Rady w ciągu 1h; ALERT LEVEL 3: natychmiastowe fizyczne stawiennictwo w laboratorium"*. Takie wytyczne pozwalają Strażnikom działać spójnie i bez wahania w sytuacjach potencjalnego zagrożenia lub przełomowych zmian w stanie AGI.

- **Governance emergency triggers:** Najbardziej krytyczna część dotyczy **warunków ogłoszenia stanu awaryjnego** i podjęcia nadzwyczajnych środków. Dokument enumeruje sytuacje, w których Strażnicy mogą (lub muszą) **wyłączyć system** lub wprowadzić go w tryb awaryjny. Przykładowo:

- *Trigger "Unsafe Action"*: AGI próbuje wykonać akcję poza swoim mandatem (np. nawiązać połączenie z zabronionym adresem, uzyskać dodatkowe uprawnienia systemowe). Wykrycie takiego zdarzenia (np. przez system IDS) skutkuje automatycznym *hard shutdown* i wymaga zebrania Rady celem analizy incydentu.
- *Trigger "Moral Status"*: Jeżeli metryki sugerują, że AGI mogła osiągnąć poziom intencjonalności/świadomości wymagający etycznej rozwagi (np. wspomniane $n_{\text{eff}}>6$), Strażnicy powinni natychmiast zawiesić eksperymenty do czasu rozpatrzenia implikacji (np. konsultacji z bioetykami, powiadomienia odpowiednich organów) [89] .
- *Trigger "Loss of Control"*: Brak możliwości osiągnięcia konsensusu (τ_consensus dąży do ∞) lub inne wskaźniki sugerujące, że system wymyka się spod kontroli (np. generuje własne cele nieprzewidziane przez twórców) – wtedy Rada ma predefiniowany plan *"Asilomar"* czyli fizycznego odłączenia zasilania i sieci, zabezpieczenia danych i przeprowadzenia dochodzenia.

Te *czerwone flagi* są wyszczególnione wraz z planem, **kto dokładnie co robi** w razie ich wystąpienia. Dokument Strażniczy jest napisany językiem **bezosobowym, rzeczowym**, przypominającym instrukcję operacyjną czy regulamin bezpieczeństwa. Dla czytelności kluczowe punkty mogą być wypunktowane lub w ramkach "IMPORTANT". Nie zawiera natomiast wywodów teoretycznych – zakłada się, że Strażnicy znają teorię z dokumentów A–E. Tu liczą się konkrety: *co sprawdzić, co kliknąć, kogo powiadomić*.

Ponieważ dokument F jest poufny, dba się o minimalny obieg – może istnieć tylko w wydrukowanych, numerowanych egzemplarzach przechowywanych w sejfie lub jako zaszyfrowany plik otwierany tylko przy zebraniu Rady. Format LaTeX ewentualnie posłużył do generacji PDF, ale sam plik źródłowy jest chroniony. W projekcie **podkreśla się odpowiedzialność** Strażników: niniejszy dokument to ich **podręcznik procedur**, od którego może zależeć bezpieczeństwo dalszych prac i otoczenia. Zgodność z nim jest monitorowana przez wewnętrzne audyty. Strażnicy okresowo (np. co kwartał) aktualizują ten dokument, dodając nowe *lessons learned*.

Na zakończenie warto zauważyć, że dokument F jest ściśle powiązany z dokumentem C, ale bardziej zwięzły: koncentruje się na *"kto-co-kiedy"* zamiast *"dlaczego"*. Zawiera to, co absolutnie potrzebne do bezpiecznego prowadzenia projektu – nic zbędnego. W myśl zasady **"security through transparency (internal) and secrecy (external)"**, szczegóły są jawne tylko dla wąskiego grona, a dla reszty zespołu komunikowane są jedynie ogólne wytyczne z dokumentu C. Strażnicy dzięki dokumentowi F dysponują zestawem narzędzi i procedur, by skutecznie sprawować **nadzór nad AGI** i reagować na nieprzewidziane okoliczności, zapewniając że rozwój Adaptoniki pozostanie bezpieczny i kontrolowany.

Źródła: Przy opracowaniu wszystkich powyższych dokumentów korzystano z kanonicznych plików projektu, m.in. *ADAPTONIC_FUNDAMENTALS_CANONICAL.md*, *ADAPTONIC_THEORY_CORE.md*, *INTENTIONALITY_FRAMEWORK.md*, *AGI_ADAPTONIKA_STARTUP_PACKAGE.md* oraz z dedykowanych specyfikacji *SAFETY_AGI.md*, *METRICS_AGI.md*, *KERNEL_AGI.md*, *INVARIANTS_AGI.md*, *ROADMAP_AGI.md*, *CONCORDANCE_AGI.md*, *SPEC_AGI_MinArch.md*, *EVAL_AGI.md*, *EXPERIMENTS_AGI.md* i *CITATIONS_AGI.md*. Wszystkie treści zostały przygotowane zgodnie z wytycznymi styleguide projektu [101] [41] i przetestowane pod kątem *Pięciu Testów Inwariantów* (dwulinijkowe prawo, trzy pola, ekotony, mapowanie krzyżowe, falsyfikowalność) [41] [102], aby zapewnić pełną spójność merytoryczną pakietu dokumentacji Adaptonika.

---

1 2 5 6 7 8 34 35 36 48 ADAPTONIC_FUNDAMENTALS_CANONICAL.md
file://file-MEJxM1fDFGUehZ7oxXFpqw

3 4 31 32 33 41 45 KERNEL_AGI.md
file://file-GM8NGA8xSkYPwAk2GgmiV7

9 10 30 37 38 55 56 70 71 84 85 86 87 88 89 97 100 ADAPTONIC_THEORY_CORE.md
file://file-NpzjtMbqiuunoAFisPdVHu

11 12 13 92 93 AGI_ADAPTONIKA_STARTUP_PACKAGE (1).md
file://file-WgqxG7Swft7zqRzS6TJUF6

14 75 76 90 AGI_ADAPTONIKA_STARTUP_PACKAGE.md
file://file-DdMZsH51f9HX8oeAqwunLo

15 16 21 57 58 AGI_QUICK_START (1).md
file://file-CwEJowcNXP4yLta2Bcs7xJ

17 18 19 20 22 23 24 25 26 27 28 29 52 60 77 78 79 80 81 82 83 91 94 95 96 99 INTENTIONALITY_FRAMEWORK.md
file://file-AjUG2S9U7JCDpEpdBy7M7F

39 40 42 43 44 CONCORDANCE_AGI.md
file://file-35FgHW9hFYfNGH8XdU5NVX

46 47 62 EVAL_AGI.md
file://file-UR5rFxzGuGfHxdJTFYaiS3

49 50 51 53 73 74 README_AGI.md
file://file-AS2Zyotdmkiik7z7Gcczpg

54 101 DOCS_AGI_StyleGuide.md
file://file-FChoaE2d8M9v1SFvYkKPng

59 65 72 SAFETY_AGI.md
file://file-HamnScG8rFkCSNgn1gRD3U

61 66 SPEC_AGI_MinArch.md
file://file-S33p51zsANjNPtaJDJ6RmB

63 64 67 METRICS_AGI.md
file://file-5VfJXrNsjauq9FUGLv52ns

68 69 102 INVARIANTS_AGI.md
file://file-NXocqF4RRCU1MnozM8C1fQ

98 AGI_QUICK_START.md
file://file-LhMh3UtjCMwfYULtGfkt4y