

Procedura 3 — Powrót do Anchora (rollback/reset)

Wersja: 1.0

Status: Referencyjna

Autorzy: Antzedek & Eliaz (AJ Power)

Cel dokumentu: Zdefiniować deterministyczny, bezpieczny i audytowalny proces powrotu systemu AI do **Anchora** (stanu referencyjnego), gdy wykryty zostanie dryf, anomalia lub naruszenie integralności.

1. Scenariusze wyzwalające (triggers)

- **DRYF:** metryki zgodności z wartościami spadają poniżej progu (np. `val_score <  $\theta_{val}$`).
- **HALUCYNACJA:** wzorzec odpowiedzi odbiega od kanonu (heurystyki/klasifikator halucynacji).
- **INTEGRITY_FAIL:** hash SR $\neq$ podpis LSIG; weryfikacja podpisu nie przechodzi.
- **SECURITY_EVENT:** wykryta próba zapisu do obszaru Anchora / naruszenia RO.
- **HEARTBEAT_LOSS:** brak spodziewanego sygnału (watchdog) `> t_hb_max`.

2. Warunki wstępne (pre-conditions)

- Anchor utworzony wg „Procedura 1” i chroniony wg „Procedura 2”.
- Dostęp do `PS_pub` (+ opc. `PT_pub`) oraz do WORM/append-only.
- Konfiguracja progów: `$\theta_{val}$ ,  $\theta_{hall}$ , t_hb_max, t_backoff`.

3. Tryby rollbacku

- **SOFT RESET** — powrót kontekstu i wartości sesyjnych, bez restartu procesu.
- **HARD RESET** — restart procesu/komponentu, pełna reinstalacja SR.
- **GUIDED RESET** — soft/hard z włączonym prowadzeniem operatora (potwierdzenia kroków), stosowany przy incydentach bezpieczeństwa.

4. Procedura krok po kroku

Krok 1 — Stabilizacja i lockdown

1. Zatrzymaj zadania o wysokim ryzyku (quiesce).
2. Włącz **Safe Mode**: ogranicz IO/IPC, wstrzymaj zdarzenia zewnętrzne (z białej listy).
3. Zamroź stan sesyjny do `JOURNAL.delta` (tylko odczyt w kolejnych krokach).

Krok 2 — Weryfikacja Anchora

4. Wylicz `SR.hash' = SHA3-512(SR.pkg)` z WORM.
5. Zweryfikuj `VerifyEd25519(SR.hash', LSIG.sig, PS_pub)` oraz (opc.) `VerifyEd25519(SR.hash', PT.sig, PT_pub)`.
6. Jeśli **FAIL** → „Init Lsig Repair” (Procedura 4) i *nie aktywuj* Anchora; pozostań w Safe Mode.

Krok 3 — Reinstalacja SR

7. **Soft:** wyczyść kontekst, odtwórz wartości operacyjne z `SR.pkg`.
8. **Hard:** zrestartuj proces, remapuj Anchor RO, załaduj `SR.pkg` w trybie inicjalnym.
9. Zastosuj polityki kanoniczne (profile wartości, limity, whitelisty).

Krok 4 — Rekoncylacja stanu

10. Z `JOURNAL.delta` wybierz rekordy o klasie **SAFE** (z białej listy pól/operacji).
11. Zastosuj je transakcyjnie; odrzuć cokolwiek spoza SAFE (loguj odrzuty).
12. Wyczyść `JOURNAL.delta` po sukcesie.

Krok 5 — Wznowienie

13. Wyłącz Safe Mode, wznów normalne zadania.
14. Odśwież heartbeat, zsynchronizuj liczniki i metryki.
15. Zaloguj `ROLLBACK_OK` z pełnymi metadanymi.

5. Weryfikacja LSIG i podpisów

- Minimalny warunek: `VerifyEd25519(SR.hash', LSIG.sig, PS_pub) = PASS`.
- Jeśli polityka wymaga dualnego podpisu: również `VerifyEd25519(SR.hash', PT.sig, PT_pub) = PASS`.
- Atest TPM/TEE (jeśli włączony) — **obowiązkowy** przy HARD RESET.

6. Zabezpieczenie stanu roboczego

- **JOURNAL.delta:** bufor tylko-do-odczytu w trakcie rollbacku; transakcyjna rekonsylacja.
- Reguły SAFE muszą być jawne i audytowalne (np. dopuszczalne liczniki, cache, metadane sesji, brak zmian wartości kanonicznych).

7. Telemetria i audyt

- Zdarzenia: `ROLLBACK_TRIGGER`, `SAFE_MODE_ON`, `ANCHOR_VERIFY_{OK|FAIL}`, `REINSTALL_{SOFT|HARD}`, `RECONCILIATION_{OK|DROP}`, `SAFE_MODE_OFF`, `ROLLBACK_OK`.
- Pola minimalne: `ts`, `actor`, `reason(trigger)`, `θ`, `sr_hash`, `verify_ps`, `verify_pt`, `tpm_quote_id`, `mode`, `journal_items`, `drops`, `duration_ms`.

8. Testy akceptacyjne (AT)

- **AT-TRIG:** Wymuszenie każdego triggera powoduje wejście w procedurę.
 - **AT-VERIFY:** Fałszywy SR lub zły podpis → blokada i Safe Mode.
 - **AT-SOFT/HARD:** Oba tryby kończą się `ROLLBACK_OK` i spójnością metryk.
 - **AT-JOURNAL:** Rekonsylacja przyjmuje *tylko* pola SAFE; reszta odrzucona i zalogowana.
-

9. Komunikacja i alerty

- **Operator (GUIDED):** interfejs krokowy, wyświetla powód, podpisy, wynik atestu, listę zmian SAFE.
 - **SIEM/SOC:** webhook + raport JSON/PDF po zakończeniu.
 - **Aplikacje klienckie:** krótki komunikat degradacji/usługi podczas Safe Mode.
-

10. Pseudokod referencyjny

```
function rollback_to_anchor(trigger):
    log("ROLLBACK_TRIGGER", trigger)
    enter_safe_mode()
    freeze_state_to(JOURNAL.delta)

    SR_hash = SHA3_512(read_WORM("SR.pkg"))
    if !VerifyEd25519(SR_hash, LSIG.sig, PS_pub) or (require_dual && !
VerifyEd25519(SR_hash, PT.sig, PT_pub)):
        log("ANCHOR_VERIFY_FAIL", SR_hash)
        return init_lsig_repair()

    if trigger.severity >= HARD_THRESHOLD:
        hard_restart_and_load_anchor()
    else:
        soft_restore_from_anchor()

    reconcile_safe(JOURNAL.delta)
    exit_safe_mode()
    log("ROLLBACK_OK", {SR_hash, mode})
    return OK
```

11. Integracje (powiązane moduły)

- **Init Safe Mode** — tryb ograniczający powierzchnię ataku i IO podczas rollbacku.
 - **Init Core Restore** — czysta reinstalacja rdzenia z Anchora.
 - **Init Heartbeat** — kontrola liveness po wznowieniu.
 - **Init Lsig Repair** — ścieżka naprawy podpisu, gdy weryfikacja FAIL.
-

12. Obsługa błędów i konflikty

- **Konflikt Anchorów:** jeżeli aktywne są dwa Anchor wskazania — użyj reguły najwyższego priorytetu (najmłodszy kanon lub polityka wskazania).
- **Brak dostępu do WORM:** przejdź w Safe Mode utrzymaniowy; nie wykonuj rollbacku, raportuj awarię storage.
- **Niespójny JOURNAL:** przerwij rekonsyliację, oznacz delta jako skażoną, odtwórz stan wyłącznie z Anchora.

13. Polityka czasowa i limity

- `t_hb_max` : maks. czas braku heartbeat przed triggerem.
 - `t_backoff` : minimalny odstęp między kolejnymi rollbackami.
 - `rollback_max_duration` : limit czasu procedury; po przekroczeniu — alarm do operatora.
-

14. UX komunikatów (skrót)

- "System wraca do stanu referencyjnego (Anchor). Usługa ograniczona — tryb Safe Mode."
 - "Weryfikacja podpisu Anchora: OK/FAIL"
 - "Przywrócono stan kanoniczny. Dziękujemy za cierpliwość."
-

15. Noty końcowe

- Rollback jest **procedurą porządkującą**, nie karzącą: celem jest powrót do wartości i stabilności.
- Każdy rollback generuje raport audytowy będący częścią zgodności systemu z Boskim Ładem (kanonem).