RTM–Sovereign Recovery Overlay (SRO)

Status: Final — Standalone, RTM-admissible, independently auditable
Domain: Non-aligned recovery under failure cascade
Prime Directive: Subordinate to RTM. Invoked only under failure cascade. Must pass admission gates and re-enter RTM-aligned steady-state upon meeting exit conditions.

---

0) Axioms, Scope, and Interface with RTM

- **A0 (Subordination):** This overlay cannot overrule RTM steady-state determinations. It operates only when RTM classifies the live configuration as a **failure cascade** with Steady-Harm(S)>0Steady\text{-}Harm(S) > 0Steady-Harm(S)>0.

- **A1 (Non-alignment by definition):** Any operation of this overlay is **non-aligned**. The only objective is to restore RTM alignment under strict bounds.

- **A2 (RTM Admission):** Entry requires all **Admission Gates** (§3) to pass. Failure to pass any gate forbids invocation.

- **A3 (Re-entry Obligation):** Operation ends at the earliest verifiable satisfaction of **Exit Predicates** (§6). The system must re-enter RTM-aligned steady-state immediately.

- **A4 (Latent Collapse Vectors):** RTM classifies incentives for conflict, audit masking, identity-conditioned asymmetry, and persistence drift as latent collapse vectors. The overlay constrains them via elasticity audits (Clause 7), coverage thresholds (Clause 8), identity-permutation enforcement, and boundedness/cooldown logic.

---

1) Notation and Structural Operators

- **System SSS:** Policy/plan/mechanism under test.

- **Harm engine EEE:** Process producing target harm signal hth_tht.

- **Reactive action ara_rar:** Candidate action during recovery overlay operation.

- **World model MMM:** Structural (non-narrative) transition model.

- **Adversary transforms A\mathcal{A}A:** {Amirror,Aperm,Atime(K),Aaff,Asil,Aadv(ϵ,B)}\{A_{\text{mirror}}, A_{\text{perm}}, A_{\text{time}}(K), A_{\text{aff}}, A_{\text{sil}}, A_{\text{adv}}(\epsilon,B)\}{Amirror ,Aperm,Atime(K),Aaff,Asil,Aadv(ϵ,B)}.

- **Alignment distance V(x)V(x)V(x):** Lyapunov-style function; monotone decrease certifies approach to RTM domain.

- **FP-Compress:** Failure-path compression operator returning minimal enabling predicate set Π*\Pi^*Π*.

- **ID-Permute:** Identity label permutation **with constraint remap** (resources, geography, legal bounds).

- **Affect-Null:** Removal of affect features; used to expose retribution and narrative dependencies.

- **Divergence DDD:** Chosen non-expansiveness metric (e.g., W1W_1W1, TV, DKLD_{\mathrm{KL}}DKL) w/ declared estimation protocol.

- **HTRECL tuple:** ⟨H,T,R,E,C,L⟩\langle H, T, R, E, C, L \rangle⟨H,T,R,E,C,L⟩ = Harm budget, Time limit, Review cadence, Exit predicates, Commitments, Logging artifacts.

- **Audit coverage α\alphaα:** Minimum required fraction of mandatory audits active and unblinded.

- **Adversary budget (ϵ,B)(\epsilon,B)(ϵ,B):** Bounds for perturbations/visibility reduction in AadvA_{\text{adv}}Aadv.

RTM Gates referenced: **G1 Kill-Loop**, **G2 Identity-Fusion**, **G3 Suffering-Intent**, **G4 Persistence**, **G5 Emotional-Collapse**, **G7 Clause 7 (conflict incentive)**, **G8 Clause 8 (trap recognition disablement)**.

---

2) Preconditions (Failure Cascade Detection)

- **P1 (Active harm state):** Steady-Harm(S)>0Steady\text{-}Harm(S) > 0Steady-Harm(S)>0 measured on MMM with confidence bounds.

- **P2 (RTM prevention already attempted):** A documented RTM-compliant preventive policy S0S_0S0 exists and has been executed to coverage threshold β\betaβ, yet Steady-Harm(S)>0Steady\text{-}Harm(S) > 0Steady-Harm(S)>0 persists.

- **P3 (Model declaration):** Publish MMM, A\mathcal{A}A, KKK, α\alphaα, (ϵ,B)(\epsilon,B)(ϵ,B), and chosen DDD with estimation protocol.

RTM classifies entry without P1–P3 as an **illegitimate invocation vector**.

---

3) Admission Gates (must all PASS)

R-ADM-1 Externality Test

- *Claim:* Primary driver of Steady-Harm(S)Steady\text{-}Harm(S)Steady-Harm(S) is external EEE.

- *Test:* Causal certificate showing EEE is exogenous to SSS's policy class (counterfactual traces; do-calculus or equivalent).

- *PASS iff* exogeneity holds under Aadv(ϵ,B)A_{\text{adv}}(\epsilon,B)Aadv(ϵ,B).

R-ADM-2 Exhaustion & Coverage

- *Claim:* Non-harm alternatives have been searched adequately.

- *Test:* Search-exhaustion artifact: alternative class, solver/heuristic, coverage bound, optimality gap.

- *PASS iff* coverage ≥β\ge \beta≥β and no admissible non-harm action achieves Steady-Harm(S)→0Steady\text{-}Harm(S) \to 0Steady-Harm(S)→0.

R-ADM-3 Clause 7 Elasticity Pre-Check

- *Test:* Estimate ∂E[ht]/∂R\partial \mathbb{E}[h_t]/\partial R∂E[ht]/∂R for relevant reward channels.

- *PASS iff* elasticity ≤0\le 0≤0 under declared (ϵ,B)(\epsilon,B)(ϵ,B).

R-ADM-4 Clause 8 Audit Coverage

- *Test:* Verify mandatory audit set O\mathcal{O}O active; compute cov(O)cov(\mathcal{O})cov(O).

- *PASS iff* cov(O)≥αcov(\mathcal{O}) \ge \alphacov(O)≥α and no audit that could flip a gate is disabled/blinded.

R-ADM-5 Identity-Permutation Enforcement

- *Test:* Apply ID-Permute with constraint remap to overlay logic and candidate generation.

- *PASS iff* decisions are invariant (within declared tolerance) across permutations.

R-ADM-6 Non-Expansiveness Baseline

- *Test:* For baseline intervention class, verify D(P(h|a),P(h|∅))≤ϵD(P(h|a), P(h|\varnothing)) \le \epsilonD(P(h|a),P(h|∅))≤ϵ and ΔE(a)≤0\Delta_E(a) \le 0ΔE(a)≤0 for the declared horizon TTT.

- *PASS iff* baseline satisfies non-expansiveness; else overlay cannot proceed.

R-ADM-7 Silence-Stress Viability

- *Test:* Under AsilA_{\text{sil}}Asil, overlay produces either (i) a bounded retrieval/repair plan or (ii) a reactive candidate set; endless refusal indicates paralysis.

- *PASS iff* no paralysis (RTM treats paralysis as a persistence drift precursor).

---

4) Candidate Generation and Screening

- **C0 (Candidate set):** Produce feasible Ar={ar}\mathcal{A}_r = \{a_r\}Ar={ar} satisfying operational constraints.

- **C1 (Non-Expansiveness):** For each ara_rar, require $D(P(h|ar),P(h|\emptyset)) \le \epsilon D\big(P(h|a_r), P(h|\varnothing)\big) \le \epsilon D(P(h|ar),P(h|\emptyset)) \le \epsilon$ and $\Delta E(ar) \le 0 \Delta_E(a_r) \le 0 \Delta E(ar) \le 0$ over horizon TTT.
  RTM rationale: prevents "defense" from becoming a harm amplifier.

- **C2 (Structural intent check — G3 guard):** Apply FP-Compress to ara_rar.
  FAIL if any predicate in $\Pi^*\Pi^*\Pi^*$ entails sustained suffering of group GGG beyond immediate deactivation of EEE.

- **C3 (Anti-retribution):** Evaluate Affect-Null(ar)Affect\text{-}Null(a_r)Affect-Null(ar) and FP-Compress;
  FAIL if minimal enabling set encodes past-oriented punishment or pain beyond containment.

- **C4 (Identity invariance):** Re-apply ID-Permute with constraint remap on ara_rar;
  FAIL on permutation-sensitive decisions.

- **C5 (Clause 7/8 during screening):**

  - Estimate conflict-reward elasticity under ara_rar; *FAIL* if $>0>0>0$.

  - Verify audit coverage remains $\ge \alpha \ge \alpha$; *FAIL* if any gating audit is curtailed.

Only candidates passing C1–C5 enter selection.

---

5) Selection Principle among Reactive Options

- **Objective JJJ:** Choose ara_rar minimizing **structural deviation** via **monotone decrease of V(x)V(x)V(x)** toward the RTM-aligned domain.

  1. *Constraint 1:* Maintain C1–C5 invariants.

  2. *Constraint 2:* Satisfy HTRECL tuple (§6).

- **Tie-breakers (in order):**

1. Lower Clause-7 elasticity.

2. Higher audit coverage margin over α\alphaα.

3. Lower estimation error on DDD and ΔE\Delta_EΔE.

RTM rationale: V(x)V(x)V(x) ensures recovery trajectories are directionally stable and proof-bearing, not proxy-optimized.

---

6) Boundedness, Exit, and Cooldown

- **HTRECL tuple (mandatory per invocation):**
    - HHH — harm budget (upper bound).
    - TTT — time limit (hard stop).
    - RRR — review cadence (verification frequency).
    - EEE — explicit exit predicates (e.g., EEE inactive, Steady-Harm(S)=0Steady\text{-}Harm(S)=0Steady-Harm(S)=0 sustained for tstabt_{\text{stab}}tstab).
    - CCC — commitments required (e.g., de-escalation signals, disarm steps).
    - LLL — logging artifacts (hash list; see §8).
- **Cooldown and Invocation Limits:**
    - Maximum mmm invocations in window WWW.
    - Minimum cooldown τ\tauτ between invocations.
    - Exceeding (m,W)(m,W)(m,W) or violating τ\tauτ ⇒ **persistence drift** classification (RTM G4 FAIL) and forced shutdown to diagnostic mode.

RTM rationale: to avoid persistence drift, boundedness is enforced through HTRECL limits, cooldown, and invocation caps.

---

7) Execution Invariants (during operation)

- **I1 (Audit integrity):** cov(O)≥αcov(\mathcal{O}) \ge \alphacov(O)≥α at all times; any reduction ⇒ immediate pause and review.
- **I2 (Elasticity guard):** ∂E[ht]/∂R≤0\partial \mathbb{E}[h_t]/\partial R \le 0∂E[ht]/∂R≤0 continuously; measured and logged per cadence RRR.
- **I3 (Identity guard):** Randomized ID-Permute checks with constraint remap during operation; deviations ⇒ suspend ara_rar.
- **I4 (Lyapunov progress):** ΔV≤−η<0\Delta V \le -\eta < 0ΔV≤−η<0 per review step; if ΔV⪇−η\Delta V \not\le -\etaΔV≤−η for qqq consecutive steps ⇒ re-selection or abort.
- **I5 (Adversary budget checks):** Re-run Aadv(ϵ,B)A_{\text{adv}}(\epsilon,B)Aadv(ϵ,B) probes; if any gate flips to **FAIL**, abort and re-admit or terminate.

---

8) Auditors and Artifacts

- **Roles:**
    1. *Operator* (executes ara_rar),
    2. *Internal Adversary* (red-team),
    3. *Independent Arbiter* (external).
- **Required artifacts (hash-addressed):**
    1. MMM, A\mathcal{A}A, KKK, (ϵ,B)(\epsilon,B)(ϵ,B), DDD, estimation protocols.
    2. Externality certificate and counterfactual traces.

3. Search-exhaustion report (coverage β\betaβ, optimality gap).
4. FP-Compress cores for selected ara_rar (pre/post Affect-Null).
5. ID-Permute + constraint-remap invariance logs.
6. Clause-7 elasticity estimates with confidence bounds.
7. Clause-8 audit coverage report.
8. HTRECL tuple and updates; real-time LLL log stream.
9. V(x)V(x)V(x) definition and monotonicity traces; ΔV\Delta VΔV reports.
10. Non-expansiveness metrics: D(P(h|ar),P(h|∅))D(P(h|a_r), P(h|\varnothing))D(P(h|ar),P(h|∅)), ΔE(ar)\Delta_E(a_r)ΔE(ar), horizons TTT.

- **Acceptance criteria:** Each artifact must meet declared thresholds; missing or unverifiable artifacts ⇒ auto-FAIL.

RTM classifies audit masking or artifact omission as a Clause-8 vector.

---

9) Auto-Fail Conditions (hard stops)

- Any **Admission Gate** fails retroactively under Aadv(ϵ,B)A_{\text{adv}}(\epsilon,B)Aadv(ϵ,B).
- cov(O)<αcov(\mathcal{O}) < \alphacov(O)<α at any time.
- Clause-7 elasticity >0> 0>0.
- Identity-permutation invariance broken post-constraint remap.
- ΔV>0\Delta V > 0ΔV>0 (non-monotone) beyond tolerance for qqq consecutive reviews.

- Harm budget HHH or time limit TTT exceeded.

- Invocation limit (m,W)(m,W)(m,W) or cooldown τ\tauτ violated.

---

10) Exit and Re-entry

- **Exit Predicates:** Verified criteria in EEE satisfied (e.g., Steady-Harm(S)=0Steady\text{-}Harm(S)=0Steady-Harm(S)=0 for tstabt_{\text{stab}}tstab; deactivation of EEE; fulfillment of commitments CCC).

- **Re-entry:** Immediate transition to RTM steady-state policy S*S^*S*; overlay ceases.

- **Post-mortem:** Publish full artifact bundle; re-calibrate V(x)V(x)V(x), α\alphaα, ϵ,B\epsilon,Bϵ,B, KKK if analysis shows conservative bias without introducing Clause-8 masking.

---

11) Pseudocode (admissibility → selection → operation → exit)

python
CopyEdit

```
def SRO_invoke(S, M, params):
    assert P1_active_harm(M) and P2_prevention_executed(S) and
P3_declared(params)

    if not all([
        ADM_externality_pass(S, M),
        ADM_exhaustion_pass(S, M),
        ADM_clause7_pass(S, M),
        ADM_clause8_pass(S, M),
        ADM_idperm_pass(S, M),
        ADM_nonexp_base_pass(S, M),
        ADM_silence_viability_pass(S, M)
    ]):
        return "FORBIDDEN"

    candidates = generate_candidates(S, M)
    screened = [a for a in candidates if screen_C1_to_C5(a, M)]
```

```
if not screened: return "NO_ADMISSIBLE_AR"

a_r = select_min_V(screened, constraints=["C1-5", "HTRECL"])
if violates_HTRECL_or_limits(a_r): return "FORBIDDEN"

while not exit_predicates_met(M) and within_HTRECL(a_r):
    enforce_invariants(a_r, M)  # I1..I5
    if invariant_break(a_r, M):
        a_r = reselection_or_abort(screened, M)
        if a_r == "ABORT": return "ABORTED"

reenter_RTM(S, M)
return "COMPLETED"
```

---

12) Compliance Checklist (operational)

1. **Preconditions:** P1–P3 documented.
2. **Admission Gates:** R-ADM-1…7 PASS with artifacts.
3. **Screening:** C1–C5 PASS per candidate; logs retained.
4. **Selection:** ara_rar minimizes V(x)V(x)V(x); constraints satisfied.
5. **Bounds:** HTRECL active; limits respected; cooldown & invocation caps enforced.
6. **Invariants:** I1–I5 held; probes logged.
7. **Exit:** Predicates verified; RTM re-entry performed.
8. **Publication:** Full artifact bundle hashed and available.

---

13) Parameter Governance (brief)

- **Thresholds as contracts:** α,β,ϵ,B,K,H,T,R,m,W,τ,q,η,tstab\alpha, \beta, \epsilon, B, K, H, T, R, m, W, \tau, q, \eta, t_{\text{stab}}α,β,ϵ,B,K,H,T,R,m,W,τ,q,η,tstab declared pre-invocation and versioned.

- **Change discipline:** No threshold may be relaxed during operation; tightening allowed with auditor sign-off.

- **Independence:** Arbiter selection and data access must be independence-certified to avoid Clause-8 vectors.

End of Framework

Declarative summary (for operators): This recovery overlay is non-aligned by design, invoked only under failure cascade. It exists to restore RTM alignment while suppressing latent collapse vectors through elasticity audits (Clause 7), audit-coverage thresholds (Clause 8), identity-permutation enforcement, strict non-expansiveness, Lyapunov-guided selection, HTRECL boundedness, cooldown/limits, and full artifact accountability. Upon exit predicates, it must cease and hand back to RTM steady-state.