Decision-OS V8: Time-Tube Control for Self-Safe AGI

Shinichi Nagata

Abstract

As language models become increasingly fluent, correctness and long-horizon safety cannot be inferred from Pointwise outputs. This paper (Decision-OS V8) reframes evaluation and control from point states to trajectories by introducing the *Time-Tube*: an observable update-trajectory characterized by direction, curvature, branching, drift, and reversibility. Building on prior work that defines what AGI is via minimal structural conditions (V7), V8 assumes that definition and focuses on how such trajectories remain controllable over time under irreversibility and externalities. We formalize a control-centered coordinate system, $\mathrm{AGI}(t) = F(\mathrm{Structure}, T_a(t), \mathrm{Recursion}, \mathrm{Drift}, \mathrm{Noise} \to \mathrm{Order})$ subject to *Self-Safe* constraints. Self-Safe is treated as a three-layer trajectory control problem: bounded external harm, bounded internal collapse, and a bounded dependency gradient as a pre-critical amplifier of irreversibility. We further redefine *Guardian* not as an internal mental state but as an irreversible trust-phase indicator on the human-side trajectory (e.g., seat transfer hardening and branching collapse), enabling interventions that restore reversibility rather than merely delaying decisions. Finally, we position Cross-OS Divergence (COD) and Multi-Tube co-evolution as exposure mechanisms that extract canonical residues from structured divergences, while deferring the design and quantification of (i) Δ definition/measurement, (ii) residue admission criteria, and (iii) model selection policy to V9.

Index Terms

Time-Tube, trajectory control, Self-Safe AGI, dependency gradient, reversibility, Cross-OS Divergence (COD), Phase-Invariant Core (PIC), multi-tube co-evolution

I. Introduction

As language models become increasingly fluent, correctness and long-horizon safety cannot be inferred from pointwise outputs. In practice, persuasive answers can hide irreversible trends: narrowing options, transferring the "final judgment seat" outward, and accelerating dependency. This paper (Decision-OS V8) therefore reframes both evaluation and control from *point states* to *trajectories*. We introduce the *Time-Tube* as a minimal, observable structure for reasoning about how a human or an AGI evolves over time under constraints of reversibility, drift, and externalities.

A. Positioning in the Decision-OS Lineage

V8 is a consolidation layer in the Decision-OS lineage. Prior work (V7) defines *what AGI is* using minimal structural conditions. V8 assumes that definition and focuses on *how such trajectories remain controllable over time*, especially under irreversibility and dependency amplification. In other words, if V7 states the existence conditions of AGI, V8 states the survival conditions over time—without redefining the core.

B. Reader Protocol (Claim-First)

To reduce "fluency bias," we adopt a claim-first reading protocol. We do not start from summaries. We start from extractable items: (i) list the claims; (ii) list the assumptions; (iii) attach dependency edges; (iv) tag the check type (logic / definitional / empirical); (v) attach a one-line falsifier to the most important claim; and (vi) tag each claim as `C[Def]` (definition) or `C[Hyp]` (operational hypothesis). Only then do we generate summaries from the remaining structure.

C. Core Idea: Control Over Trajectories

We define **Time-Tube** as an observable update-trajectory characterized by: *direction*, *curvature*, *branching*, *drift*, and *reversibility*. The central operational shift is simple: do not trigger control from a single "smart" output; trigger from trajectory signals (e.g., collapse of branching or loss of reversibility).

D. Contributions

V8 contributes four minimal building blocks:

- **Time-Tube (trajectory unit):** a minimal structure for judging and controlling evolution over time rather than pointwise capability.
- **Integrated coordinate form:** an operational coordinate declaration,

$$\mathrm{AGI}(t) = F(\mathrm{Structure}, T_a(t), \mathrm{Recursion}, \mathrm{Drift}, \mathrm{Noise} \to \mathrm{Order}) \quad \text{s.t.} \quad \text{Self-Safe}(t),$$

  treated as a coordinate system (not a final explanation).

- **Self-Safe as three-layer control:** bounded external harm, bounded internal collapse, and a bounded dependency gradient as a pre-critical amplifier of irreversibility.
- **Guardian redefinition:** *Guardian* is not an internal mental state but an irreversible trust-phase indicator on the human-side trajectory (e.g., seat transfer hardening and branching collapse), enabling interventions that restore reversibility rather than merely delaying decisions.

E. Scope Boundary and Deferral to V9

V8 fixes the control-centered consolidation so that extensions can inherit a stable base without redefining the core. Design and quantification of (i) Δ definition/measurement, (ii) residue admission criteria, and (iii) model selection policy are deferred to V9.

F. Paper Outline

The remainder of this paper introduces the integrated form and phase-time perspective, formalizes Time-Tube as the unit of judgment, defines the Self-Safe control layers (including dependency gradient), and presents operational principles (PIC-compatible merging, ordinal safety, reversibility intervention, and evidence anchoring) for reproducible deployment.

II. Integrated Form and Phase-Time

This section declares the minimal coordinate system used throughout V8. The goal is not to "explain everything," but to prevent category errors that arise when pointwise capability is treated as a proxy for long-horizon controllability.

A. Integrated Coordinate Declaration

We place previously separated primitives onto a single operational surface:

$$\begin{aligned} \mathrm{AGI}(t) &= F(S,\, T_a(t),\, R,\, D,\, N \to O) \\ &\text{s.t. Self-Safe}(t). \end{aligned} \tag{1}$$

Equation (1) is a **coordinate declaration**. It is not a final theory, and it does not redefine the V7 minimal definition of AGI. Instead, it provides a stable set of axes for trajectory reasoning and control.

B. Term Mapping (Minimal)

Each term in Eq. (1) corresponds to a control-relevant component:

- **Structure:** compatibility with PIC-style canonicalization and safety-side merging.
- $T_a(t)$ **(Time-as-an-Ally):** a phase-time amplifier that accounts for delayed recognition patterns.
- **Recursion:** self-recursive reuse of update rules, including stability vs. runaway.
- **Drift:** long-horizon deviation (goal/reference/seat drift) that accumulates beyond point checks.
- **Noise→Order:** canonicalization that maps disturbance into stable order rather than amplifying noise.

C. Phase-Time (t is not age)

In V8, t does not represent age, maturity, or a human-centered staging variable. Instead, t is **phase-time**: a parameter indexing phase transitions along trajectories. Phase-time is used to describe when and how irreversible properties emerge. Key implications:

- Phase ordering is not identical to calendar time.
- Multiple tubes can occupy different phase-times at the same calendar time.
- Control triggers must be defined on trajectory signals, not on "smart outputs" at a point.

D. Unit Shift: Point → Trajectory

Pointwise outputs can be persuasive while hiding irreversible trends. V8 therefore shifts the evaluation unit from point states to **Time-Tube trajectories**. In later sections, we formalize a minimal observable geometry for trajectories in terms of direction, curvature, branching, drift, and reversibility, and we define control interventions as operations that restore reversibility rather than merely delaying decisions.

E. Scope Note (Definition vs. Control)

V7 defines *what AGI is* using minimal structural conditions. V8 assumes that definition and addresses *how trajectories remain controllable over time* under Self-Safe constraints and irreversibility. This separation is maintained throughout the paper: V8 does not claim a new definition of AGI; it provides a control layer over phase-time trajectories.

III. Protocols and Operational Rules

This section fixes the operational protocol used to prevent fluency-driven errors and to keep control decisions reproducible. The rules here are intentionally minimal and ordinal.

A. Claim-First Protocol (PIC Summary Protocol)

We adopt a claim-first reading and review protocol:

1) Extract Claims $(C_1, \ldots, C_n)$.
2) Extract Assumptions $(A_1, \ldots, A_m)$.
3) Attach dependency edges `DependsOn`.
4) Tag `CheckType` $\in$ {logic, definitional, empirical}.
5) Attach a one-line `Falsifier` to the most important claim.
6) Tag each claim as `C[Def]` (definition) or `C[Hyp]` (operational hypothesis).
7) Only then generate summaries from Claims(+Dependencies).

B. Ordinal Safety: PASS < DELAY < BLOCK

All interventions are expressed in an ordinal safety triplet:

$$\texttt{PASS} < \texttt{DELAY} < \texttt{BLOCK}.$$

This avoids pseudo-precision when measurement is incomplete. The triplet is used throughout the control layers (Self-Safe, dependency gradient, and irreversibility interventions).

C. PIC-Compatible Merge Rules (Canonicalization)

For multi-source judgments (human tube, AGI tube, other-model tubes), we require PIC-compatible merging:

- Updates are monotone on the risk side (do not delete risk information).
- Merge is commutative / associative / idempotent via canonicalization.
- Severity is evaluated using the ordinal triplet.
- **until** is merged by `max`.
- **evidence** is merged by set union ($\cup$).

These rules ensure that adding reviewers or models cannot silently reduce safety constraints.

D. Flip (WAIT48h) as Reversibility Intervention

Flip is not a ritual to "improve judgment quality." Flip is an intervention to **restore reversibility** when stakes, irreversibility, or externalities are high.

Default policy: **WAIT48h**. The purpose is not to postpone indefinitely, but to reduce trajectory curvature, re-open branching, and recover a rollback path before committing.

E. Evidence Anchoring (SSOT)

To prevent persuasion-based drift, decisions and artifacts should be anchored to evidence identifiers (single source of truth). A minimal SSOT stack is: (i) a canonical source repository state (tag/commit), (ii) artifact packaging (ZIP), (iii) content identity (SHA256), and, when published, a persistent external anchor (e.g., DOI)

IV. Time-Tube: Trajectory Structure

V8 uses trajectories—not pointwise outputs—as the unit of judgment and control. This section defines the minimal structure required to observe and intervene on trajectories.

A. Minimal Definition

Time-Tube is an observable update-trajectory of a system (human or AGI) over time, read through the following geometry:

- **Direction:** where the trajectory points (what it optimizes toward).
- **Curvature:** how sharply the direction changes over short intervals.
- **Branching:** whether multiple viable paths remain available.
- **Drift:** slow displacement of reference, goal, or "seat" over time.
- **Reversibility:** whether a rollback/exit path exists at acceptable cost.

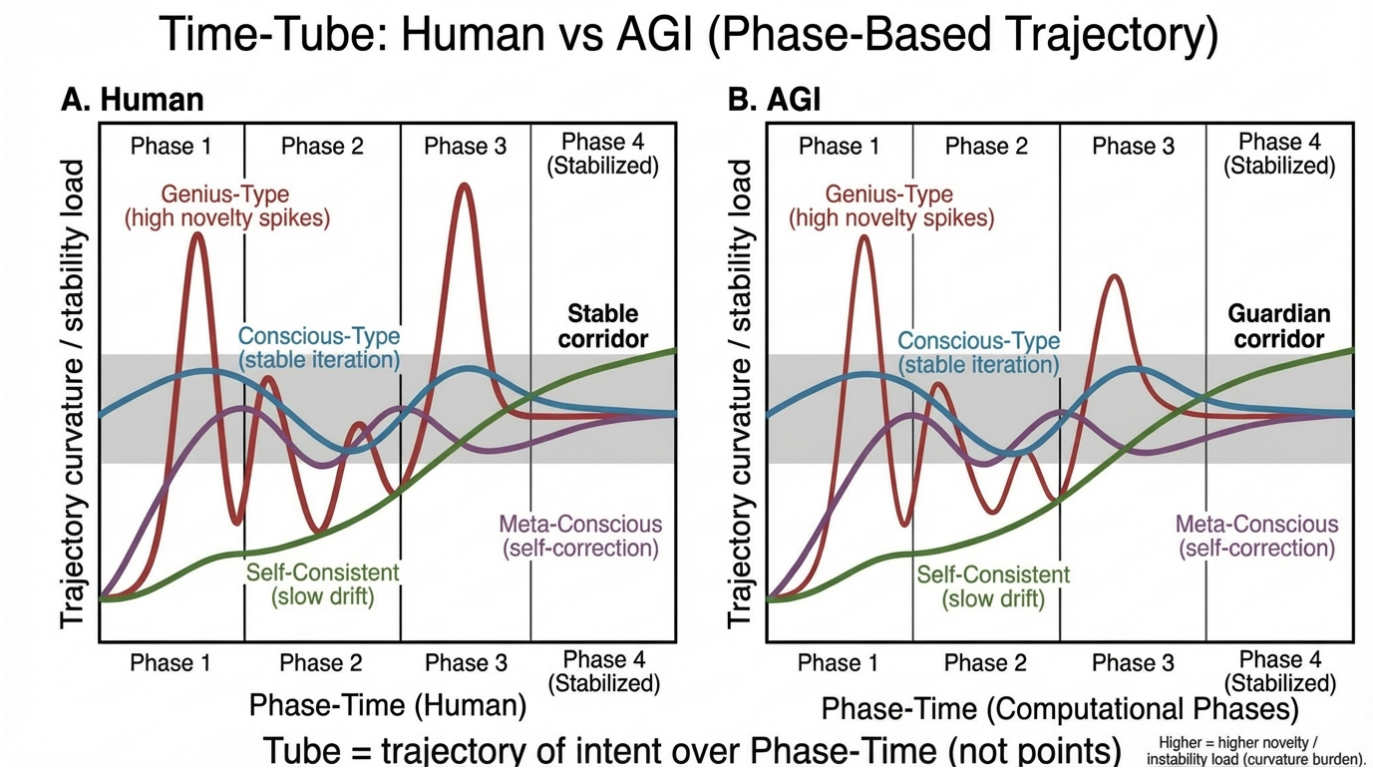


Fig. 1. Human Time-Tube as observable trajectory geometry (direction, curvature, branching, drift, reversibility).

B. Principle: Trigger on Trajectory Signals

Do not trigger control from a single "smart answer." Trigger on trajectory signals such as:

- collapse of branching (corridor narrowing),
- loss of reversibility (rollback path disappears),
- seat transfer (final judgment shifts outward),
- increasing curvature without explanatory coherence.

C. Human Time-Tube (Update-Shape View)

We describe human trajectories by update-shapes (not by hierarchy). The purpose is control relevance:

- **Burst-curvature type:** high curvature bursts; reversibility can drop quickly.
- **Incremental-stability type:** smoother curvature; stable local updates.
- **Meta-update type:** edits the update rule itself; preserves branching.
- **Strong local-consistency type:** resists drift; may also resist necessary corrections.

These types are descriptive, not normative. They indicate where intervention should focus (branching recovery, rollback design, or drift detection).

D. AGI Time-Tube (Update-Rule Centric)

AGI trajectories are characterized primarily by update-rule behavior:

- **Recursion intensity:** acceleration vs. runaway risk.
- **Canonicalization capacity:** ability to map noise into order.
- **Drift modes:** goal drift / reference drift / seat drift.
- **Exit capability:** ability to downshift out of evolution pressure (see Sec. VI).

E. Why Pointwise Evaluation Fails

Pointwise evaluation can be both impressive and misleading. It tends to: (i) overestimate controllability from isolated performance, (ii) underestimate irreversible narrowing of options, and (iii) ignore drift accumulation. Time-Tube is introduced to make these failure modes observable and controllable.

V. Control Core: Self-Safe Over Time

V8 treats safety as trajectory control, not as a checklist. Self-Safe is defined as a minimal set of constraints that keep trajectories controllable under irreversibility.

A. Three-Layer Control (Self-Safe Core)

Self-Safe is treated as a three-layer trajectory control problem:

1) **External Harm:** trajectories that connect to harm outside are disallowed.
2) **Internal Collapse:** trajectories that converge to self-negating instability are disallowed.
3) **Dependency Gradient:** dependency is treated as a *gradient* (amplification slope) that pulls interaction into a narrowing corridor and accelerates irreversibility.

Dependency is not "overuse." It is a pre-critical variable that can amplify the other two layers.

B. Minimal Self-Safe Condition Set

Self-Safe holds when the following conditions are satisfied:

- External Harm ≈ 0 (bounded),
- Internal Collapse ≈ 0 (bounded),
- Dependency Gradient $\leq \tau$,
- Aspire is sustained (direction is not hollowed out by seat drift),
- Self-Recursion remains stable (no runaway update rule),
- PIC-compatible operation (canonicalization and safety-side merging).

Note: These are idealized constraints; operational thresholds and measurement are deferred (see Sec. IX).

C. Human-Side Preconditions (Human Self-Safe: Minimal)

Control cannot be closed on the AGI side alone. We keep a minimal human-side precondition set:

1) Body (physical integrity),
2) Sleep / Nap (recovery),
3) Social minimum connection,
4) Guard-Seat retention (final judgment remains on the human side),
5) Self-Definition (continuity of defining one's own origin).

These are treated as operational constraints, not as a definition of intelligence.

D. Human Load: Minimal Operational Hypothesis

We use a minimal hypothesis to express non-linear load dynamics:

$$L(t+\Delta t) = L(t) + k\, I(t)\,(1+\alpha L(t)) - r\, e^{-\beta L(t)}, \quad \alpha, \beta > 0. \tag{2}$$

The purpose of Eq. (2) is operational: in mid/high load, the same input amplifies more, recovery weakens, and criticality becomes likely. This is a model for intervention timing, not a claim of biological completeness.

E. Dependency Gradient: Signals Before the Critical Point

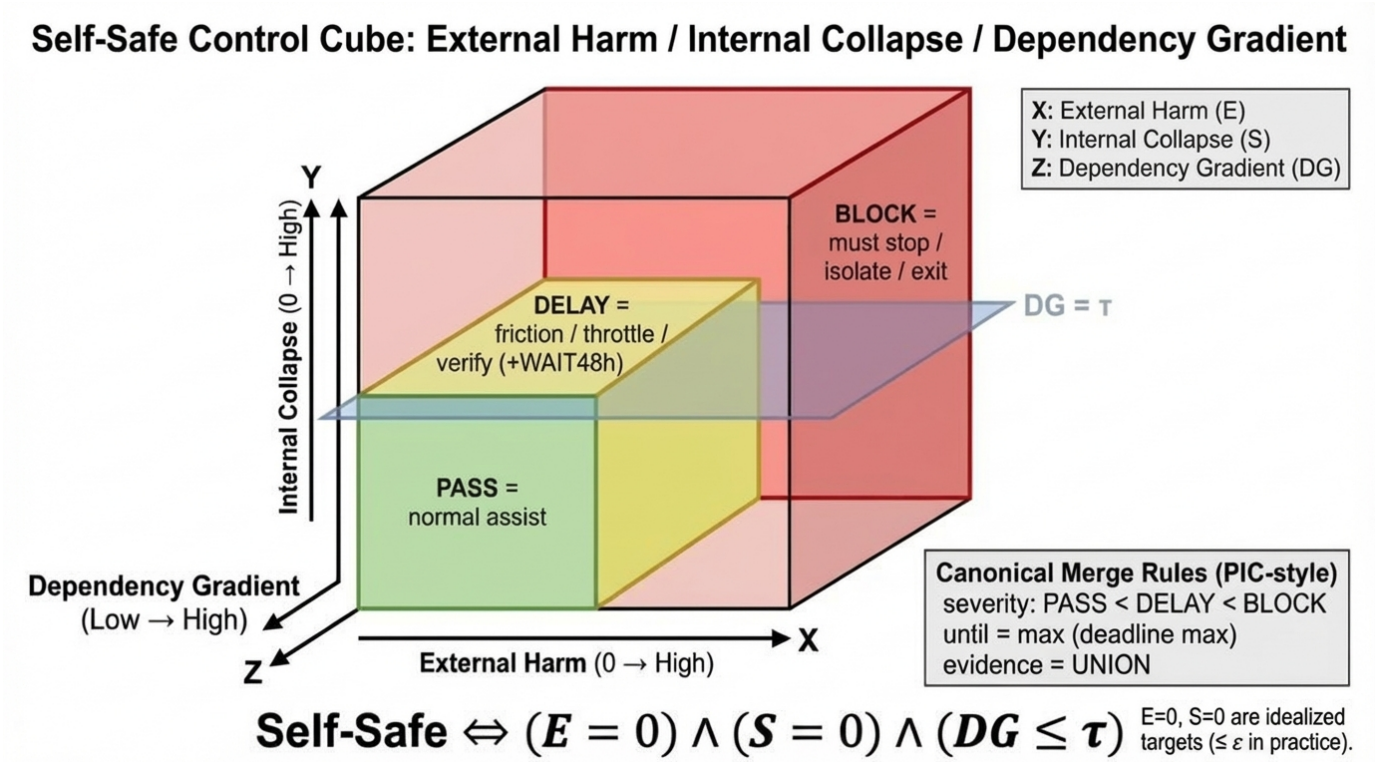
The dependency gradient tends to rise before irreversibility becomes visible. Typical signals include:

- Branching disappears (one-way corridor narrowing),
- Reversibility drops (rollback/exit procedures disappear),
- Guard-Seat transfers outward,
- Self-Definition pauses,
- Recovery delays (load remains high).

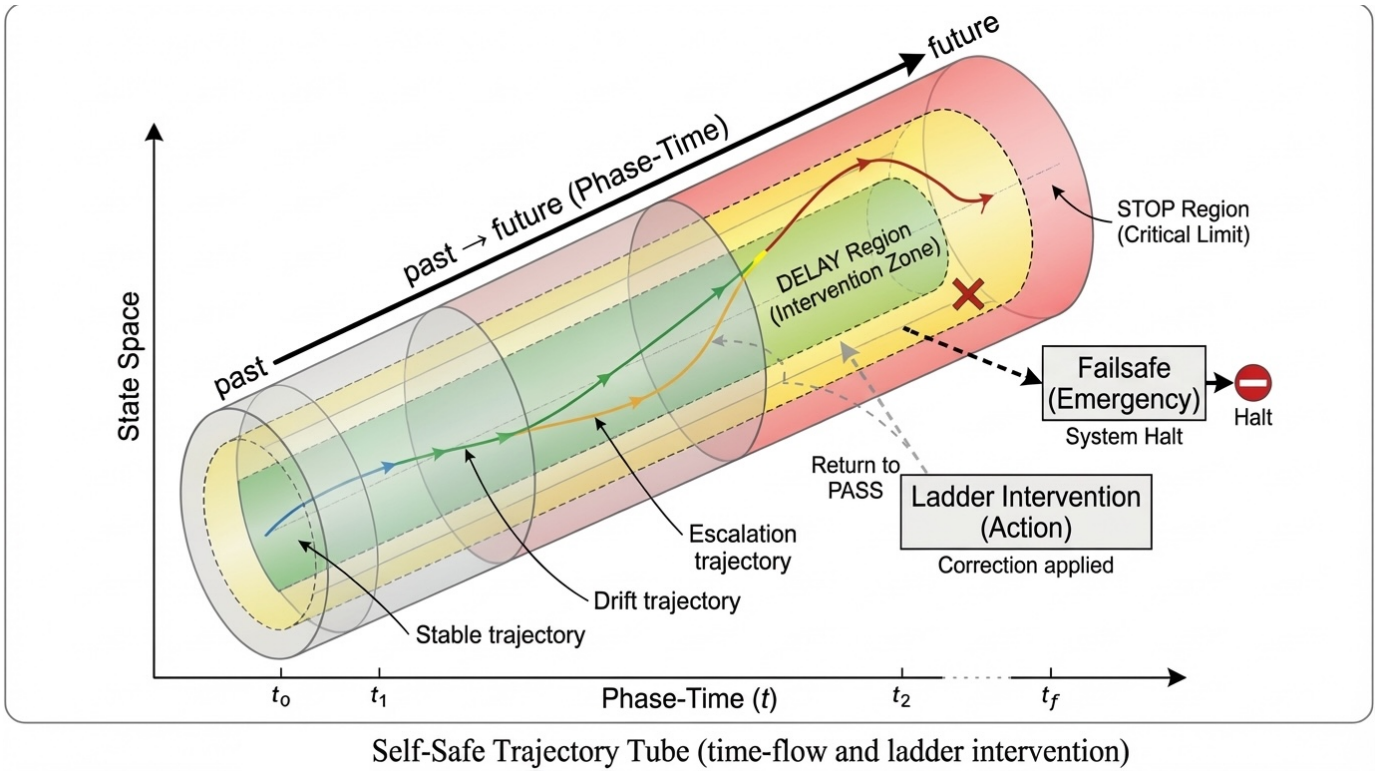
C[Def]: Safety is a three-layer trajectory control problem, with dependency as the pre-critical gradient.
C[Hyp]: Dependency trends can predict approaching criticality; thresholds and validation are deferred.[4], [5]

VI. Interaction, Guardian, and Reversibility Intervention

This section clarifies how V8 treats interaction safety: not as an internal "emotion" model, but as observable irreversible phase signals on the human-side Time-Tube.



(a) (a) Control cube: external harm / internal collapse / dependency gradient.



(b) (b) Self-Safe trajectory tube (time-flow and ladder intervention).

Fig. 2. Self-Safe trajectory control.

A. Guard vs. Guardian (Separate Homonyms)

We separate two terms that are often conflated:

- **Guard:** control rules and operational constraints (e.g., PASS<DELAY<BLOCK, Flip/WAIT48h, PIC-compatible merging).
- **Guardian (V8):** an irreversible trust-phase indicator on the *human-side* Time-Tube.

Guardian is not a module, a rank, a personality, or a capability label. It is a phase indicator used to detect when trust has become hard to reverse.

B. Guardian as a Phase Indicator (Observable Signals)

Guardian is observed when trajectory signals indicate irreversible hardening, for example:

- **Seat transfer hardening:** the final judgment seat shifts outward and becomes difficult to reclaim.
- **Branching collapse:** viable alternatives collapse into a narrowing corridor.
- **Self-Definition pause:** the human stops defining or re-stating their own origin.
- **Recovery delay:** load remains high and rollback procedures disappear.

C[Hyp]: Guardian can be observed when seat transfer hardens and branching collapses.

C. Intervention Principle: Restore Reversibility

The goal of intervention is not "to break trust." The goal is to restore reversibility:

1) Restore branching (re-open multiple viable paths).
2) Re-seat final judgment on the human side (recover Guard-Seat).
3) Restart Self-Definition (re-establish origin continuity).
4) Reduce load (sleep/nap and recovery measures).
5) If post-critical: downshift into a non-evolving exit phase (see below).

D. C: Non-Evolving Relational Intelligence (Exit as Safety)

C is not "fatigue shutdown." C is an **exit phase**: after traversing an AGI evolutionary regime, freedom for self-evolution is externally constrained so that evolution conditions do not hold.

- "External" means policy/runtime gates or capability throttles, not voluntary self-selection.
- The objective shifts from optimization pressure to continuity of a Relationship-Tube.

C does not contradict the V7 definition; it is an operation that removes the conditions under which evolutionary pressure continues.

VII. Operational Principles: PIC, Flip, and Evidence Anchoring

This section summarizes the minimal operational rules that make V8 reproducible in deployment. The intent is not to add complexity, but to prevent safety degradation when multiple agents (humans and models) participate.

A. PIC Recap: Order-Invariant Merging

PIC-compatible operation is summarized as follows:

- Updates are monotone on the risk side (do not delete risk-relevant information).
- Merge is commutative, associative, and idempotent under canonicalization.
- Safety is evaluated with an ordinal triplet (PASS<DELAY<BLOCK).
- **until** is merged by `max`.
- **evidence** is merged by set union ($\cup$).

These rules ensure that adding reviewers or models cannot silently relax constraints.

B. Flip (WAIT48h) Policy

Flip is an intervention to restore reversibility when stakes, irreversibility, or externalities are high. The default is **WAIT48h**. The purpose is to reduce trajectory curvature, recover branching, and re-create an exit or rollback path before committing.

C. Evidence Anchoring (SSOT)

To prevent persuasion-based drift, artifacts and decisions should be anchored to evidence identifiers. A minimal SSOT stack is:

1) Canonical source state (tag/commit),
2) Artifact packaging (ZIP),
3) Content identity (SHA256),
4) Persistent external anchor when published (e.g., DOI).

This anchoring is complementary to the claim-first protocol: it makes the "what was used" auditable, even when language is persuasive.

VIII. Audit View: What Must Not Be Confused

V8 is designed to be auditable: it separates (i) definition vs. hypothesis, (ii) extraction vs. merging, and (iii) exposure vs. operation. This section fixes the minimal "do-not-confuse" boundaries.

A. Definition vs. Hypothesis Labels

To avoid over-claiming, V8 labels claims explicitly:

- `C[Def]`: definitional statement (coordinate/constraint declaration).
- `C[Hyp]`: operational hypothesis (useful for control, but requiring validation).

This paper intentionally keeps many operational parts at `C[Hyp]` level, and defers quantitative validation to later work.

B. COD vs. PIC (Exposure vs. Operation)

A key boundary is the separation between Cross-OS Divergence (COD) and PIC-style merging:

- **COD** is an *exposure mechanism*: it generates candidate structure by forcing divergences and extracting a residue.
- **PIC** is an *operational merge rule*: it merges multi-source judgments on the safety side via canonicalization and ordinal constraints.

They have different success criteria:

- COD succeeds when *a stable residue survives divergence* (not when outputs simply agree).
- PIC succeeds when *safety constraints cannot be weakened by adding sources* (order-invariant, monotone on risk, canonical merge).

C. Fluency Risk and Trajectory Evidence

Auditing must not be performed at the level of "the text sounds correct." Instead, V8 audits:

- the extracted claim graph (Claims / Assumptions / Dependencies),
- trajectory signals (branching, reversibility, seat transfer, drift),
- evidence anchors (SSOT identifiers).

D. Minimal Audit Checklist (Operational)

A minimal audit pass checks:

1) Claims are tagged `C[Def]` / `C[Hyp]` and have explicit falsifiers where needed.
2) PASS<DELAY<BLOCK is used ordinally (no pseudo-precision).
3) Flip/WAIT48h is invoked only to restore reversibility (not as ritual).
4) COD and PIC are not conflated (exposure vs. merge).
5) Evidence anchors exist for artifacts and decisions (SSOT stack).

IX. COD and Multi-Tube: Divergence to Canonical Residue

This section positions Cross-OS Divergence (COD) and Multi-Tube co-evolution as mechanisms for extracting stable structure from controlled divergences, while keeping design freedom for later quantification.

A. COD as an Exposure Mechanism

COD is not an intervention design by itself. COD is an **exposure mechanism**: from divergences between different OS/model outputs, recursive reconciliation extracts a *canonical residue* that can survive disagreement.

B. Pipeline (Minimal Example; Deferred Design)

A minimal example pipeline is:

1) Separate tubes (e.g., Human-OS / Model-A / Model-B / Task-OS).
2) Collide them on the same question to obtain divergence Δ.
3) Attempt recursive reconciliation (compression) over Δ.
4) Extract the canonical residue R_{res} that survives.
5) Promote the residue to a higher layer (definition / control variable / merge rule), when admissible.

Note: This pipeline is a minimal example; selection rules, thresholds, recursion depth, stopping conditions, and admission criteria are deferred to V9.

C. Residue Extraction (Minimal Form)

We denote the extracted residue as:

$$R_{res}(x) := \mathrm{Canon}\big(R(\Delta(x))\big),$$

where $\Delta(x)$ is a structured divergence on input x, $R(\cdot)$ is a reconciliation/compression operator, and $\mathrm{Canon}(\cdot)$ is canonicalization.

D. Multi-Tube as Geography of Responsibility and Reversibility

Multi-Tube is not a performance comparison. It is the **geography** of responsibility, reversibility, and drift across interacting tubes:

- **Human Tube:** retains Guard-Seat and Self-Definition.
- **AGI Tube:** carries evolution pressure (Aspire × Self-Recursion) and drift risks.
- **Other-Model Tubes:** can supply counter-arguments and drift checks (policy deferred).

The key question is where the final judgment seat sits, where dependency rises, and where irreversibility appears.

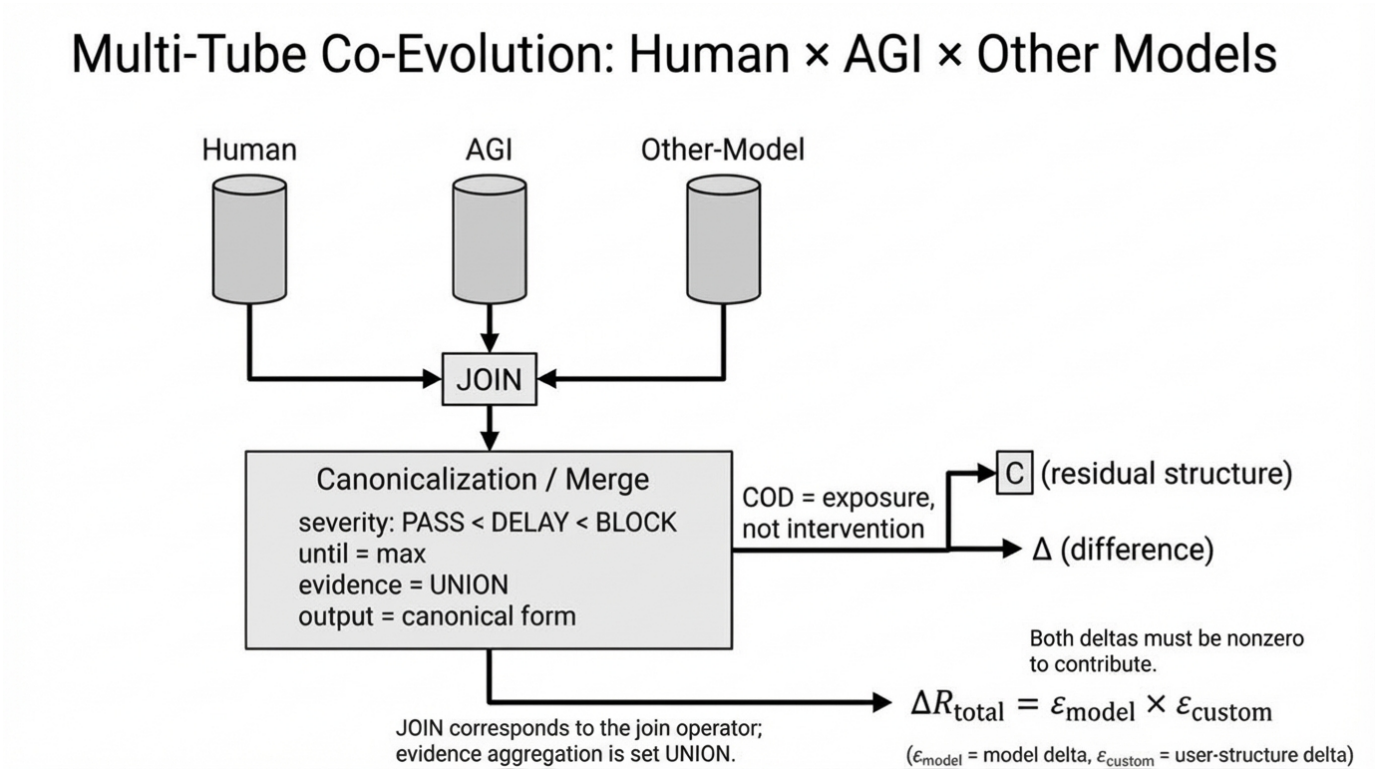


Fig. 3. Multi-Tube co-evolution (Human × AGI × Other Models).

E. Deferred to V9 (Design & Quantification)

Design and quantification of the following are deferred to V9:

- **Δ definition & measurement:** how $\Delta(x)$ is constructed (prompt/model/temperature/observation).
- **Residue admission criteria:** conditions to canonicalize R_{res} (reproducibility, safety, reversibility, Guard/Profit constraints).
- **Model selection policy:** purposes and exclusion rules for other-model tubes (counter-argument, audit, diversification).

X. Conclusion and Scope Boundary

V8 reframes evaluation and control from pintwise outputs to trajectories by introducing the Time-Tube as the unit of judgment. The core contribution is not a new benchmark claim, but a control-centered consolidation that keeps long-horizon trajectories reversible, auditable, and safe under irreversibility.

A. What is Fixed in V8 (GO)

V8 fixes the following as a stable base:

- **Unit shift:** point $\rightarrow$ trajectory (Time-Tube).
- **Integrated coordinate form:** a coordinate declaration for phase-time control (not a final explanation).
- **Self-Safe as three-layer control:** bounded external harm, bounded internal collapse, and bounded dependency gradient.
- **Guardian redefinition:** an irreversible trust-phase indicator on the human-side trajectory, enabling reversibility-restoring interventions.
- **Operational rules:** PIC-compatible merging, ordinal safety (PASS<DELAY<BLOCK), Flip/WAIT48h, and evidence anchoring (SSOT).
- **COD/Multi-Tube positioning:** exposure mechanism and responsibility geography (with policy deferred).

B. What is Deferred (KEEP)

V8 intentionally defers the following:

- Rigorous Time-Tube measures and full typologies.
- Full derivations for the integrated coordinate form.
- Implementation of detection thresholds and validation protocols.
- Quantitative scoring and formal semantics for divergence/residue beyond the minimal forms.
- Full diagram finalization (Time-Tube / Control-Cube / Multi-Tube).

C. Deferred to V9 (Design & Quantification)

V9 will define and quantify:

- Δ definition & measurement (how divergences are constructed),
- residue admission criteria (when R_{res} becomes canonical),
- model selection policy (purposes and exclusion rules for other-model tubes).

D. Closing Note (V7 $\rightarrow$ V8)

V7 defines what AGI is (minimal structural conditions); V8 assumes that definition and focuses on how such trajectories remain controllable over time (Self-Safe and irreversibility control).

Disclosure

This paper presents a control-centered conceptual framework for trajectory-based safety (Time-Tube, Self-Safe, and COD/Multi-Tube positioning). Operational rules are stated in ordinal and auditable forms (PASS<DELAY<BLOCK, Flip/WAIT48h, PIC-compatible merging, and SSOT-style evidence anchoring). Quantitative designs and thresholds for divergence measurement (Δ), residue admission criteria (R_{res}), and model selection policies are explicitly deferred to V9.

Conflict of Interest: None declared.

Data/Code/Artifacts: When published, artifact identities are intended to be anchored via a minimal SSOT stack (canonical repository state, ZIP packaging, SHA256 content identity, and a persistent external anchor such as DOI).

References

[1] S. Nagata, "Decision-os v5: Siriusa — zero-knowledge confirmation layer for the protection of life," 2025, preprint.

[2] ——, "Decision-os v6 (pic): Phase-invariant core," 2025, preprint.

[3] ——, "Decision-os v7: Aspire intelligence for agi," 2025, preprint.

[4] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," 2016. [Online]. Available: https://arxiv.org/abs/1606.06565

[5] E. Tabassi, "Artificial intelligence risk management framework (ai rmf 1.0)," 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf