Decision-OS V9: Impact-Weighted Release

Continuous-Pass Release under Public Tube — Time V2 × MMAR × Hindsight Control

Shinichi Nagata

Abstract

Decision-OS V9 defines a release procedure for high-impact systems operating under public repetition. The core problem is not a lack of intelligence, but failure amplification caused by (i) hindsight overwrite, (ii) responsibility laundering, and (iii) repeated exposure with independence breakdown (the *Public Tube*). V9 addresses this by fixing *As-of* (temporal fidelity), *Seat* (decision rights and accountability), and *Release* (impact-weighted continuous pass). Time V2 separates evaluation into a Playback-only lane bound to an As-of Pack and a Forward-only update lane expressed as explicit deltas, preventing post-hoc narrative edits of the past. For public deployment, V9 treats repetition as a first-class risk factor: release is granted only after sustained passing over an observation window that scales with impact and exposure. We also describe interaction primitives (CAP and DFR) and a disagreement-based method (MMAR) to extract invariants without collapsing uncertainty into persuasive coherence. Overall, V9 shifts the criterion from "smart answers" to procedures that remain stable under audit, hindsight, and repeated public trials.

Index Terms

Impact-weighted release, temporal fidelity (As-of), accountability (Seat), Public Tube, continuous pass, Time V2, CAP/DFR, MMAR

I. Introduction

Safety is not the ability to "get the right answer." Safety is the ability to *stop*, *delay*, and *roll back* when the world flips.

Decision-OS V8 formalized this stance as *Time-Tube control*: safety is enforced over a line (trajectory), not by claiming one-shot correctness at a point. However, real failures are often amplified not by a single mistake, but by how evaluation and responsibility are handled *after the fact*. Time does not only amplify value; through narrativization and hindsight, it can amplify exemption, condemnation, and the erosion of accountability.

Decision-OS V9 does not aim to predict the future. V9 provides an operational procedure that is harder to break: it fixes the as-of moment when a decision was made, separates improvements into a future delta lane, and defines *Release* as an impact-weighted *continuous pass* rather than one-shot correctness.

A. Lineage: from "Decision" to "Impact-Weighted Release"

The series position can be summarized as follows:

- V4: Decision as form (an entry OS for seeing structure).
- V5: Protection-of-life gates (stop / verify / confirmation).
- V6: PIC (canonicalization; phase-invariant merge).
- V7: Aspire Intelligence (recursion-aware operational framing).
- V8: Time-Tube Control (point → trajectory; intervention over time).
- V9: Impact-Weighted Release (Time V2; Public Tube + Hindsight Control).

V8 focuses on internal safety (intervention over a trajectory). V9 canonicalizes this into external operational rules that remain valid under public release and repeated exposure.

B. As-of fixation and update-lane separation

V9 introduces *As-of Fixation*: a decision is evaluated against the bundle of information, constraints, and objectives available at the time—the *As-of Pack*.

V9 also introduces *Update Lane Separation*:

- **Playback-only lane:** explanations are constrained to the As-of Pack.
- **Forward-only update lane:** improvements are proposed only as future operational deltas.

This separation prevents hindsight from "beating the past" and destroying learning. It also prevents responsibility laundering via post-hoc reframing, where time is used to justify exemption or to weaponize condemnation.

C. Impact-weighted release under public repetition

V9 defines *Release* as a gate-based condition, not a narrative justification. Release is granted only after *sustained passing*, with longer observation windows for higher-impact decisions. This accumulates institutional defense cost against the *Public Tube* phenomenon: even small success probabilities can eventually "hit" through repetition, large n, and independence breakdown.

In short, V9 shifts the center of gravity from "smart answers" to procedures that do not break. Safety is not intelligence—it is a gate.

D. Minimal structure: three outcomes and three fixations

V9 commits to a minimal structure that all subsequent chapters extend without breaking.

a) Outcomes (judgment lane):

- **PASS** (proceed).
- **DELAY** (do not decide yet; recheck window applies).
- **BLOCK** (do not proceed).

b) Fixations (non-negotiable anchors):

- **As-of:** fix temporal fidelity (As-of Pack).
- **Seat:** fix responsibility and decision rights (Decision Owner / Auditor).
- **Release:** publish only via impact-proportional observation windows and continuous pass.

E. Contributions

This paper contributes an operational constitution for high-stakes decisions under public release:

- **Time V2 procedure:** As-of Pack → Playback-only → Forward-only update (fixed protocol).
- **Hindsight control:** review as temporal-fidelity preservation plus exactly one future delta (no character judgment).
- **Public Tube framing:** repeated exposure and independence breakdown as first-class risk factors for release.
- **Impact-weighted continuous pass:** release scales with impact and requires sustained passing, not one-shot success.

This motivates Time V2 as a procedural defense against hindsight bias [1].

II. Overview

This chapter summarizes the core claim of Decision-OS V9:

> Release is not a narrative. Release is a procedure. The procedure must remain valid under hindsight, repetition, and public exposure.

For public-facing systems, documenting intended use, limits, and evaluation context is part of safe release [2].

V9 defines the minimal operational structure that survives these pressures by fixing (i) temporal fidelity (*As-of*), (ii) responsibility (*Seat*), and (iii) publication conditions (*Release*).

A. Problem: why "good reasoning" still fails in public

Even when a decision is made with careful reasoning, public reality introduces three failure amplifiers:

- **Hindsight overwrite:** post-hoc narratives replace the original constraints and objectives, destroying learning.
- **Responsibility laundering:** explanations drift into blame or exemption, moving responsibility away from the decision owner.
- **Repetition amplification:** small probabilities "hit" under large n, especially when independence breaks.

These amplifiers are not solved by smarter answers. They are solved by constraining what can change, and where it is allowed to change.

B. V9 in one page: As-of, Seat, Release

V9 is built from three anchors.

a) (1) As-of: temporal fidelity as a hard constraint: A decision must be reviewed against the information and constraints available at the time of execution. V9 packages this as an *As-of Pack* and binds explanations to *Playback-only*. Improvements are not allowed to "edit the past"; they must be expressed as *forward-only deltas*.

b) (2) Seat: responsibility and decision rights are fixed before evaluation: V9 treats responsibility assignment as a pre-evaluation gate. A review must first determine who holds the *decision seat* (Decision Owner) and who audits (Auditor). Without this, evaluation becomes either scapegoating or absolution, both of which destroy operational learning.

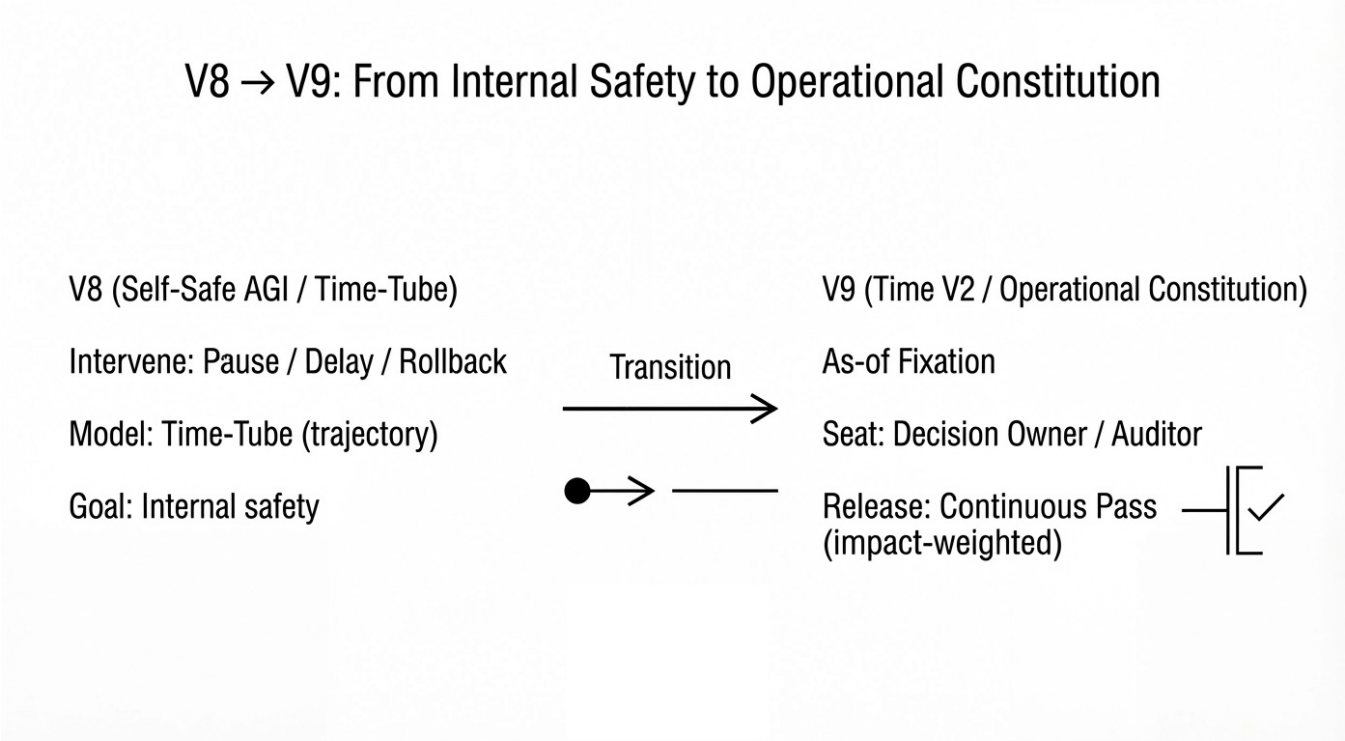

Fig. 1. From V8 (trajectory intervention) to V9 (impact-weighted release): As-of fixation, Seat fixation, and continuous-pass Release under Public Tube.

c) (3) Release: impact-weighted continuous pass: V9 defines release as an operational gate that scales with impact. Higher impact requires longer observation windows and repeated passing. Release is not granted by one-shot success.

C. Public Tube: repeated exposure as a first-class risk factor

V9 treats public release as a *tube* that accumulates attempts, viewers, and opportunities for misuse. Even when a single attempt has a low probability of harm, repeated exposure increases aggregate risk. In addition, independence breaks in public: copying, re-sharing, and coordinated trials make outcomes more correlated. Figure 2 illustrates this accumulation and union effect.

Therefore, release must account for:

- repeated trials (n growth),
- independence breakdown (correlation),
- and impact magnitude (harm scale).

D. Outputs: three outcomes plus explicit recheck

V9 standardizes judgment into three outcomes:

- **PASS:** proceed.
- **DELAY:** do not decide yet; set an explicit *until* time and recheck conditions.
- **BLOCK:** do not proceed.

DELAY is not indecision. It is a controlled safety state that preserves agency while refusing premature commitment.

E. Roadmap of the paper

The remainder of this paper proceeds as follows:

- Section III defines the fixed protocol (As-of Pack, CAP/DFR, release gates).
- Section IV describes the system architecture and canonical merge constraints.
- Section V details Time V2 and the anti-hindsight mechanism.
- Section VI covers interaction patterns and MMAR-style structure extraction.
- Section VII describes multi-party confirmation and Seat control under pressure.

Figure 2.

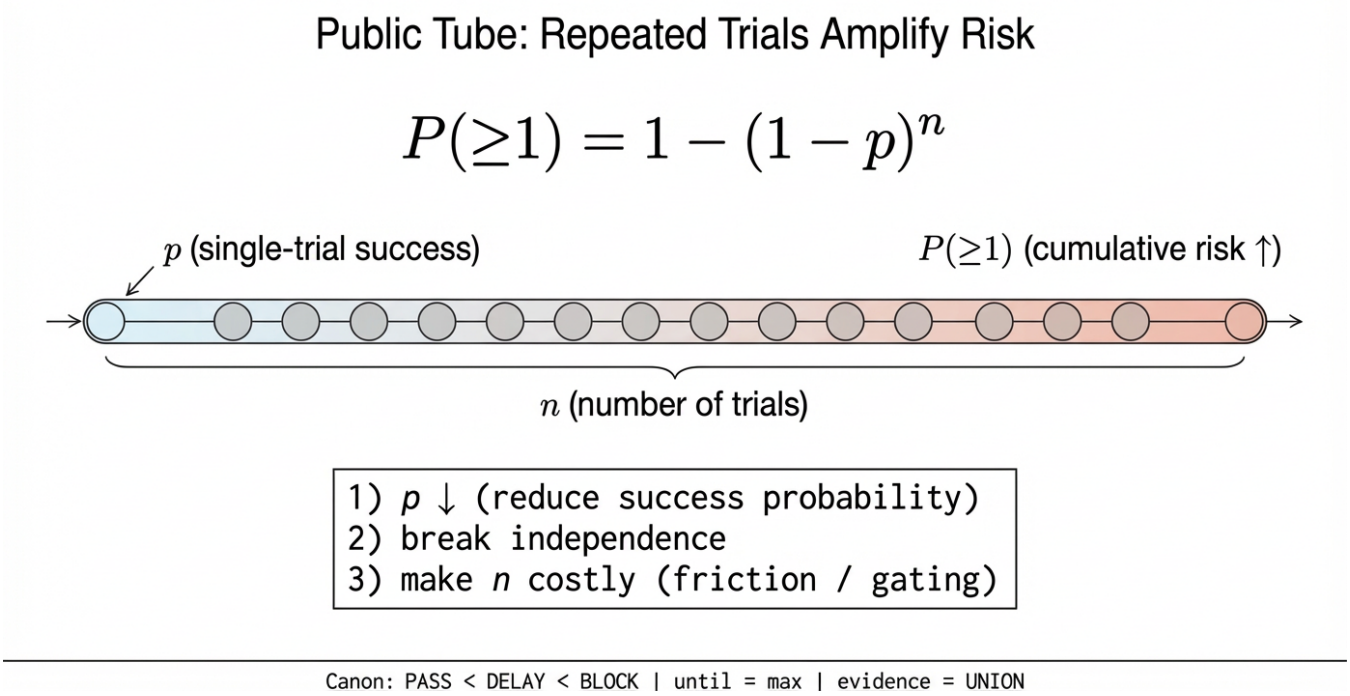

Fig. 2. Public Tube: risk accumulation under repeated exposure (growing n) and partial independence breakdown.

III. Protocol

A. DFR v1: As-of -> Delta -> Residue -> Next step

This chapter defines the fixed operational protocol of Decision-OS V9. The protocol is designed to survive (i) hindsight pressure, (ii) responsibility drift, and (iii) repeated public exposure.

V9 is intentionally procedural: it constrains *what* may change and *where* it may change. All evaluations are bound to an As-of Pack, and all improvements are expressed only as forward deltas.

B. As-of Pack: what is fixed

An *As-of Pack* is the minimal bundle that must be fixed before any evaluation:

- **Timestamp (As-of):** when the decision was executed.
- **Context:** constraints, resources, environment, and capability ceiling at the time.
- **Objective:** the objective function used at the time (including safety constraints).
- **Action:** what was actually done (not what was intended).
- **Evidence:** the observable inputs available at the time (logs, screenshots, records).

V9 binds explanations to the As-of Pack (*Playback-only*). Hindsight is permitted only as *forward-only deltas* (Section III-E).

C. Pre-evaluation gates: Seat, deadline, probability

Before judging correctness or quality, V9 fixes three gates:

a) Seat (responsibility): Assign roles explicitly:

- **Decision Owner:** holds the decision rights and responsibility for the action.
- **Auditor:** reviews the action and proposes deltas; does not take the seat.

b) Time horizon (deadline): If a decision includes a deadline or time sensitivity, record the horizon explicitly (e.g., hours, days, weeks). This prevents "deadline drift" in post-hoc narratives.

c) Probability (uncertainty): If the decision implicitly assumes a probability p, record whether p was (i) estimated, (ii) bounded, or (iii) unknown. Unknown is a valid state; it is handled via **DELAY** (Section III-G).

D. CAP v1: Claim, Dependencies, Alert-if, Recheck, Hedge

V9 uses a minimal claim protocol to prevent post-hoc story edits.

a) CAP record: Each decision claim is expressed as:

- **Claim:** the actionable statement (what is asserted).
- **Dependencies:** which assumptions must hold.
- **Alert-if:** observable signals that invalidate the claim.
- **Recheck:** when and how the claim will be revisited.
- **Hedge:** the safety action if the claim weakens (e.g., stop, downshift, rollback).

b) Constraint: A CAP record must be written in As-of terms. New information cannot be injected into the Claim retroactively; it must appear as a forward delta.

E. DFR v1: As-of -> Delta -> Residue -> Next step

V9 standardizes review into a forward-only form. This prevents two failure modes: (i) moral condemnation that destroys learning, and (ii) hindsight absolution that launders responsibility.

a) DFR steps:

1) **As-of Playback:** restate the decision under the As-of Pack (no new facts).
2) **Delta:** specify exactly what would be changed going forward (procedure, threshold, check, guard).
3) **Residue:** specify what remains valid (invariant structure that should not be "edited away").
4) **Next step:** choose one operational update to implement next (a single change, not a redesign).

b) One-change rule: A single review produces at most one operational update. If more changes are desired, they must be queued as separate deltas with separate rechecks.

F. Judgment lane: PASS, DELAY, BLOCK

V9 constrains outcomes to three states:

- **PASS:** proceed under the current protocol.
- **DELAY:** do not decide yet; set an explicit *until* time and recheck triggers.
- **BLOCK:** do not proceed; the action is rejected under current constraints.

These states apply both to private decisions and to public release decisions. The difference is that public release is additionally constrained by the Release gate (Section III-H).

G. DELAY: controlled non-commitment with explicit until

DELAY is not indecision; it is a safety state that preserves agency.

A DELAY record must specify:

- **until:** the earliest time the decision may be revisited,
- **recheck conditions:** what must be observed to revisit,
- **hedge while waiting:** what safety posture is enforced during the delay.

If a decision is high impact and uncertainty is not bounded, DELAY is the default safe state.

H. Release gate: impact-weighted continuous pass

V9 treats release as a gate that scales with impact and exposure.

We operationalize impact as a minimal product of externality, responsibility, and probability: $I(a) := E(a) \cdot R(a) \cdot P(a)$.

a) Impact weighting: Let I denote impact magnitude (harm scale if misused, or irreversible cost if wrong). Higher I requires longer observation windows and stricter passing conditions.

b) Continuous pass: Release is granted only if the system remains in PASS across a required observation window. In operational terms: the process must pass *continuously*, not once.

c) Public Tube pressure: Public release increases attempt count n and often breaks independence via copying and coordination. Therefore, release must consider:

- growth of n over time,
- correlation (independence breakdown),
- and worst-case misuse paths.

When impact is high or independence is likely to break, the protocol must shift toward DELAY or BLOCK unless strong guards are present.

We align the release gate with established risk management practice for trustworthy AI [3].

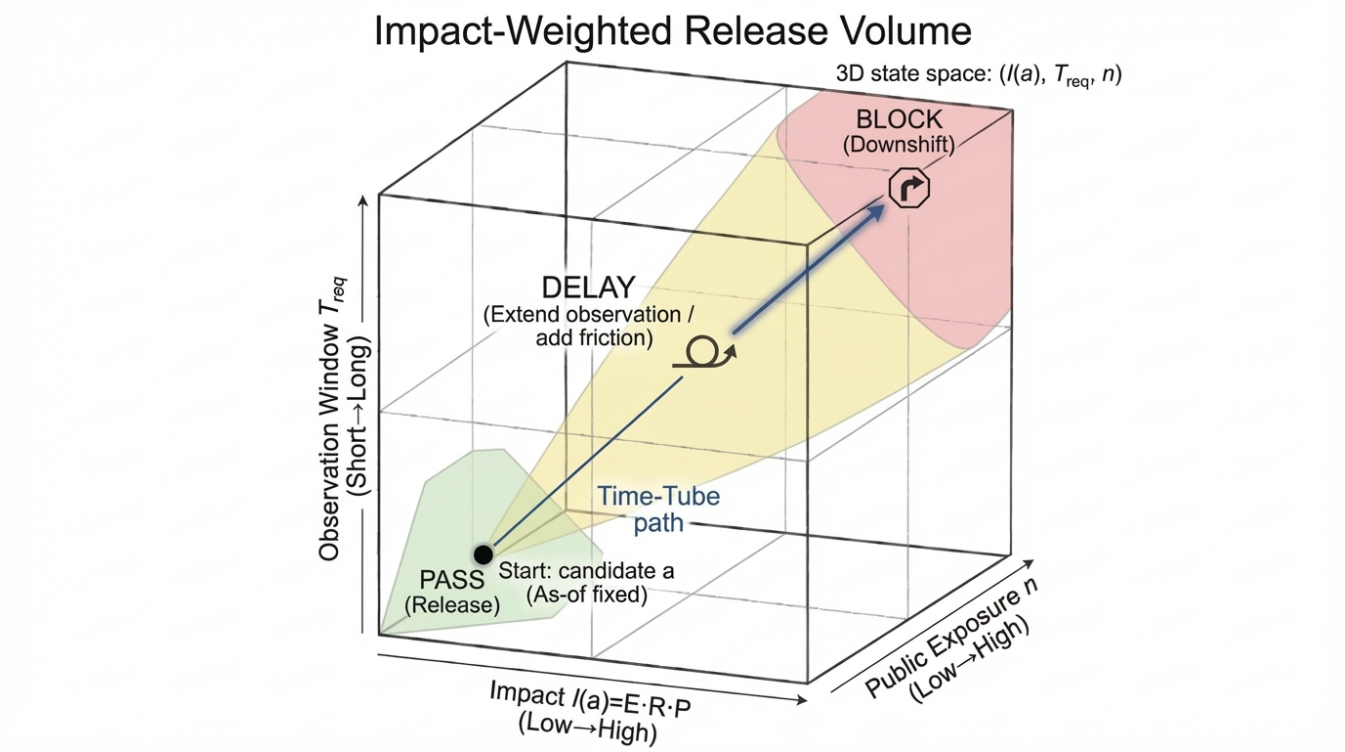

Fig. 3. Impact-weighted continuous-pass release: required release volume increases with impact and Public Tube pressure.

I. Protocol summary

The V9 protocol can be executed as the following fixed sequence:

1) Fix the As-of Pack (Section III-B).
2) Apply pre-evaluation gates: Seat, deadline, probability (Section III-C).
3) Write CAP for the claim (Section III-D).
4) Choose outcome: PASS / DELAY / BLOCK (Section III-F).
5) If reviewing: run DFR and implement exactly one forward delta (Section III-E).
6) If publishing: apply the impact-weighted continuous pass Release gate (Section III-H).

IV. Architecture

This chapter describes the system architecture that makes the V9 protocol operational. The goal is not to build a large framework, but to define a minimal structure that remains stable under iteration, public release, and audit.

V9 models decision-making as state updates. Each module emits a *delta* that is merged into a global state via a canonical merge rule. This prevents order-dependent behavior and reduces accidental "story edits".

A. Global state and module deltas

Let S be the global operational state. Each module produces an update ΔS_i. The system merges updates into a new state:

$$S' = \mathrm{Canon}\big(S \sqcup \Delta S_1 \sqcup \Delta S_2 \sqcup \cdots\big).$$

The merge operator $\sqcup$ is designed to be:

- **commutative** (order does not matter),
- **associative** (grouping does not matter),
- **idempotent** (reapplying the same delta changes nothing).

Canon$(\cdot)$ enforces a canonical form (normalization) so the same content yields the same state representation, making audits tractable.

B. Core registries

V9 maintains a small set of registries, each append-only by default:

- **As-of Pack Registry:** records the fixed bundle for each decision instance.
- **Seat Registry:** assigns Decision Owner and Auditor before evaluation.
- **CAP Registry:** stores Claim, Dependencies, Alert-if, Recheck, Hedge.
- **DFR Review Log:** stores As-of Playback, Delta, Residue, Next step.
- **Release Gate Log:** records impact weighting, observation windows, and pass continuity.
- **Public Tube Signals:** tracks exposure growth and independence breakdown indicators.

The primary design rule is separation:

- explanation is bound to As-of (Playback-only),
- improvement is written only as a forward delta (Forward-only update).

C. Severity lattice and canonical merge rules

V9 reduces control to a three-level severity lattice:

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

When multiple modules propose outcomes, the merged outcome is the maximum severity. This ensures safety dominance under disagreement.

V9 also standardizes merges for common fields:

a) Until time: DELAY uses an explicit *until*. When multiple until values exist, the merged until is the maximum time, ensuring the longest required waiting window is respected.

b) Evidence: Evidence is merged as a set union. No evidence item is deleted by later explanations; removal requires an explicit audit action.

c) Risk and impact: Risk estimates and impact levels are merged conservatively. If two modules disagree, the higher risk or higher impact dominates the merged state.

D. Canonicalization: preventing story edits

The canonicalization function Canon$(\cdot)$ enforces invariants that protect temporal fidelity and responsibility assignment:

- **As-of invariance:** As-of Pack fields cannot be modified by review outputs.
- **Seat invariance:** the decision seat cannot drift during review; Auditor cannot become Decision Owner mid-stream.
- **Delta isolation:** improvements must be written as forward deltas and cannot overwrite the As-of Playback.

This converts the common failure mode (post-hoc narrative replacement) into a constrained update form: the past is played back; only the future is edited.

E. Release pipeline as an operational gate

Release is treated as a pipeline that consumes a stabilized state and emits a release decision.

a) Inputs: The pipeline reads:

- the latest As-of Pack and Seat assignment,
- the current CAP record (including alert and recheck),
- the current severity outcome (PASS, DELAY, or BLOCK),
- impact magnitude and public exposure signals.

b) Gate rule: Release is granted only if:

- the merged outcome is PASS, and
- the continuous pass condition holds over the required observation window, scaled by impact.

If the state is DELAY or BLOCK, release is rejected by design. This keeps publication aligned with operational safety rather than narrative confidence.

F. Auditability: replay and diff

V9 supports two audit operations:

a) Replay (Playback-only): Given an As-of Pack and the recorded state deltas, an auditor can reconstruct the exact state at the time of the decision without injecting new facts.

b) Diff (Forward-only): Given a review, the auditor extracts the delta that will be applied going forward. The delta is explicit, minimal, and recorded as a single operational change.

This architecture turns evaluation from a debate into a reproducible procedure: *replay the past, then apply exactly one forward delta.*

V. Time V2 and Hindsight Control

This chapter defines Time V2: a time-aware evaluation rule that preserves temporal fidelity and prevents hindsight from rewriting the past.

Time V2 is not a forecasting tool. It is a constraint on explanations and reviews:

- the past is evaluated only under the As-of Pack (Playback-only),
- improvements are written only as forward deltas (Forward-only update).

A. Two lanes: playback-only and forward-only

Time V2 separates evaluation into two lanes.

a) Playback-only lane: The reviewer must restate the decision using only the As-of Pack. No new facts, no upgraded capabilities, no revised objectives.

b) Forward-only update lane: Any improvement is written as a delta that applies from now on. The delta must include dependencies, alert conditions, and a recheck plan.

This lane separation prevents the most common failure: a review that "wins" by importing future knowledge, and therefore destroys learning.

B. Reverse temporal evaluation: capability ceiling at the time

Time V2 requires *Reverse temporal evaluation*: past decisions are judged under the capability ceiling and constraints that existed at the time. A reviewer is not allowed to retroactively demand actions that were not feasible, not knowable, or not affordable under the As-of Pack.

Operationally:

- the As-of Pack fixes the feasible action set,
- the review may propose a future delta to expand feasibility,
- but the evaluation of the past does not borrow that expansion.

This rule prevents both condemnation by impossible standards and absolution by fabricated constraints.

C. Ta and Da: time does not increase value by itself

Time V2 separates two effects:

- $T_a(t)$: amplification that occurs only when re-evaluation under external signals confirms the structure.
- D_a: degradation or discount that captures decay, drift, or opportunity cost.

A minimal relation is:

$$V_{\text{later}} = T_a(t) \times D_a.$$

Crucially, $T_a(t)$ is not a function of time alone. It increases only through re-evaluation triggered by external signals, such as:

- independent confirmation,
- adversarial testing,
- real-world deployment evidence,
- or repeated survival under public exposure.

This prevents a common narrative abuse: claiming that time automatically validates a decision.

D. Re-evaluation triggers and recheck windows

Time V2 treats re-evaluation as a scheduled operation with explicit triggers. A claim must specify:

- **Alert-if:** signals that invalidate the claim,
- **Recheck:** when the claim will be revisited,
- **Hedge:** what safety posture holds until the recheck.

When uncertainty cannot be bounded, the default posture is DELAY with an explicit until time. Time V2 therefore turns "waiting" into a controlled safety state rather than a vague pause.

E. Misuse prohibitions

Time V2 explicitly forbids two misuses.

a) Prediction misuse: Time V2 must not be used to claim certainty about future outcomes. It is an evaluation constraint, not a prophecy engine.

b) Hindsight absolution misuse: Time V2 must not be used to erase responsibility by rewriting the past. Explanations stay within Playback-only. Improvements belong to the forward-only delta lane.

F. Operational summary

Time V2 is executed as a fixed rule set:

1) Fix the As-of Pack.
2) Run Playback-only evaluation under the capability ceiling at the time.
3) If change is needed, write exactly one forward-only delta (with alert and recheck).
4) Maintain PASS, DELAY, or BLOCK without narrative drift.

Time V2 makes reviews reproducible: the past is replayed as it was, and only the future is edited.

VI. Interaction Protocols

Decision-OS V9 treats interaction itself as a risk surface. Many failures are not caused by a lack of intelligence, but by conversational dynamics: narrative smoothing, role drift, and premature commitment.

This chapter defines interaction rules that preserve the V9 anchors: *As-of*, *Seat*, and *Release*.

A. Role fixation: Decision Owner vs. Auditor

V9 requires explicit role fixation during interaction.

- **Decision Owner** holds the decision seat and can commit an action.
- **Auditor** may challenge, test, and propose deltas, but does not take the seat.

Role swapping mid-review is prohibited. This prevents a common failure mode where responsibility drifts into "shared vibes" and learning collapses into blame or absolution.

Operationally, every interaction thread begins by stating:

- the Decision Owner,
- the Auditor,
- and whether the thread is **Playback-only** or **Forward-only update**.

B. As-of constrained prompting

When an assistant model is used, V9 constrains prompts to the As-of Pack (Section III-B). The assistant must be instructed to:

- restate the situation using only the provided As-of Pack,
- avoid importing "likely" facts not present in the pack,
- and explicitly mark unknowns as unknown (without filling gaps by narrative).

A minimal As-of prompt contract is:

> Use only the As-of Pack below. Do not add new facts. If information is missing, keep it missing. Return (i) Playback-only restatement, (ii) uncertainties, (iii) PASS/DELAY/BLOCK suggestion.

This contract prevents the assistant from producing convincing but ungrounded reconstructions that overwrite the past.

C. Conversation drift: narrative smoothing as a failure mode

Assistants often "help" by smoothing contradictions into a coherent story. In high-stakes contexts, this is a safety failure.

V9 therefore forbids three interaction behaviors in Playback-only lanes:

- **Gap filling:** inventing missing premises to complete a story.
- **Moralization:** turning review into character judgment.
- **Outcome worship:** judging past actions primarily by eventual results rather than As-of constraints.

Instead, Playback-only interaction is strictly descriptive:

- what was known,
- what was constrained,
- what was optimized,
- and what was executed.

D. Multi-Model Collision Review (MMAR)

V9 supports Multi-Model Collision Review (MMAR) as an interaction method to extract structure from disagreement.

a) Purpose: When one model produces a highly coherent narrative, it can override uncertainty. MMAR introduces controlled disagreement to surface invariants and failure modes.

b) Independence rule: Models must not share conversational memory. Each model receives the same As-of Pack, but no cross-context.

c) Minimal MMAR prompt set: For each model k, request:

- the strongest opposing argument (best counter-case),
- likely failure modes and their detection signals (Alert-if),
- and a PASS/DELAY/BLOCK recommendation with reasons tied to the As-of Pack.

d) Merge rule: MMAR outputs are merged conservatively:

- outcome severity uses max over PASS/DELAY/BLOCK,
- evidence is unioned,
- and proposed updates are queued as forward deltas (one-change rule in Section III-E).

MMAR is not used to "vote for truth". It is used to detect structural risk and to prevent narrative overwrite.

E. Interaction guardrails: stopping, delaying, rolling back

V9 makes three guard actions explicit in interaction:

a) Stop: If the action is unsafe or the seat is undefined, interaction must halt in **BLOCK** until corrected.

b) Delay: If uncertainty is high and impact is high, interaction defaults to **DELAY** with an explicit until time and recheck triggers (Section III-G).

c) Rollback: If a decision has already been executed and new evidence appears, the interaction must:

1) replay the original As-of Pack (Playback-only),
2) then propose exactly one forward delta (Forward-only update),
3) and specify a hedge action during the transition (CAP Hedge).

These guardrails keep interaction procedural rather than persuasive.

F. Output contracts: CAP + DFR as interaction primitives

To prevent drift, V9 treats CAP and DFR not as documentation, but as *interaction primitives*.

a) CAP as commitment boundary: Before committing to an action, interaction must be able to state: Claim, Dependencies, Alert-if, Recheck, Hedge (Section III-D).

b) DFR as review boundary: After an action, interaction must produce: As-of playback, Delta, Residue, Next step (Section III-E).

By enforcing these outputs, V9 converts conversation into reproducible state transitions.

G. Summary

V9 interaction is constrained by:

- **Role fixation** (Seat does not drift),
- **As-of constrained prompting** (Playback-only is factual),
- **MMAR** for structure extraction (disagreement reveals invariants),
- **Guardrails** (Stop / Delay / Rollback),
- and **CAP + DFR** as mandatory interaction outputs.

The goal is not to sound convincing. The goal is to remain correct *as a procedure* under public pressure.

VII. Multi-Party Confirmation and Seat Control

High-impact decisions often fail not because the decision owner is irrational, but because confirmation pathways are weak. Under pressure, a single actor can be nudged into irreversible actions by narrative momentum, interface friction, or social coercion.

This chapter defines a minimal multi-party confirmation layer that preserves the V9 anchors: *Seat* (decision rights), *As-of* (temporal fidelity), and *Release* (impact-weighted gating).

A. Why multi-party confirmation is needed

Single-party confirmation is fragile under:

- high-stakes urgency (deadline compression),
- asymmetric information (the user cannot verify claims),
- and repeated persuasion attempts (Public Tube pressure).

Multi-party confirmation introduces a deliberate speed bump that:

- increases detection probability for scams and coercion,
- reduces impulsive irreversible actions,
- and creates an auditable trail for later DFR updates.

B. Seat control: decision rights are not negotiable mid-stream

V9 treats Seat as a pre-evaluation gate. Seat determines who can commit actions and who can audit them.

- The **Decision Owner** may execute actions.
- The **Auditor** may challenge and propose deltas.

During high-impact flows, V9 strengthens Seat control by requiring external confirmation before execution. This prevents role drift and reduces the chance that persuasion converts into commitment.

C. Confirmation states

V9 defines a minimal confirmation state machine:

- **CONFIRM-0** (Solo): Decision Owner confirms without external party.
- **CONFIRM-1** (Buddy): One external confirmer is required.
- **CONFIRM-2** (Family / Team): Two confirmers are required for irreversible actions.

The required level is impact-weighted: higher impact requires higher confirmation level.

D. What must be confirmed

Confirmation is not "approval of intent." It is a structured check of the As-of Pack and the CAP record.
A confirmer must verify:

- **Action:** what will be executed (exactly).
- **Destination:** who receives value or access (recipient, address, endpoint).
- **Irreversibility:** whether rollback is possible.
- **CAP fields:** Dependencies, Alert-if, Recheck, Hedge.

If any of these cannot be verified, the default state is **DELAY**.

E. Speed control: delay as a protective valve

A key purpose of multi-party confirmation is to restore time. V9 treats time not as a luxury but as a safety resource.
For high-impact actions, confirmation enforces:

- an explicit *until* window (cooldown),
- a recheck step (independent verification),
- and a hedge posture while waiting (freeze, partial rollback, or no-action).

This converts "wait" into a controlled protocol rather than hesitation.

F. Audit trail: evidence union and replayability

Multi-party confirmation must be auditable.

a) Evidence union: All confirmers append evidence. Evidence is merged as a set union and cannot be deleted by later narrative explanations.

b) Replay: A later audit can replay:

- the As-of Pack,
- the CAP record,
- the confirmation state and confirmations,
- and the executed action.

This supports DFR: the past is replayed as it was; only the future is edited.

G. Failure modes and defenses

Multi-party confirmation is not a guarantee. V9 therefore models common failure modes:

- **Collusion:** confirmers are not independent.
- **Capture:** confirmers are pressured or socially coerced.
- **Rubber-stamp:** confirmation becomes a ritual with no verification.

Defenses are procedural:

- diversify confirmers when possible (independence),
- require at least one external verification step (recheck),
- and treat uncertainty as DELAY, not as permission to proceed.

H. Summary

V9 multi-party confirmation is a minimal protection layer:

- impact-weighted confirmation states (CONFIRM-0/1/2),
- structured verification of action, destination, irreversibility, and CAP fields,
- enforced cooldown and recheck windows (DELAY as protective valve),
- and an append-only audit trail (evidence union, replayability).

This layer preserves Seat control under pressure and reduces irreversible mistakes under public repetition.

Disclosure

a) *Scope and intent.* This work proposes an operational decision protocol (Decision-OS V9) for preserving temporal fidelity, responsibility assignment, and safe release under public repetition. It is not a prediction system and does not claim empirical guarantees.

b) *Safety posture.* The protocol reduces harm by enforcing procedural constraints (PASS/DELAY/BLOCK, As-of Pack fixation, Seat fixation, and impact-weighted continuous pass). If an action is high-impact, irreversible, or cannot be verified under the As-of Pack, the recommended default posture is **DELAY** or **BLOCK**.

c) *Limitations.* This paper is conceptual and procedural. It does not provide a complete empirical evaluation, formal proofs for all claims, or a comprehensive threat model. Real-world deployment requires domain-specific adaptation, independent auditing, and continuous re-evaluation.

d) *Ethical considerations.* The protocol aims to prevent hindsight-based responsibility laundering, scapegoating, and narrative overwrite. Reviews should focus on As-of constraints and forward deltas, not on character judgment. High-risk use cases (e.g., medical, legal, financial, and safety-critical operations) should require stricter confirmation and independent oversight.

e) *Conflicts of interest.* The author declares no external conflicts of interest.

References

[1] B. Fischhoff, "Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 1, no. 3, pp. 288–299, 1975.

[2] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," 2018. [Online]. Available: https://arxiv.org/abs/1810.03993

[3] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," U.S. Department of Commerce, National Institute of Standards and Technology, Tech. Rep. NIST AI 100-1, Jan. 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[4] S. Nagata, "Decision-OS V5: SiriusA — Zero-Knowledge Confirmation Layer for the Protection of Life," 2025, preprint.

[5] ——, "Decision-OS V6 (PIC): Phase-Invariant Core," 2025, preprint.

[6] ——, "Decision-OS V8: Time-Tube Control for Self-Safe AGI," 2026, preprint.