DFW v6.0 — Temporal Safety Architecture

A Long-Horizon Risk and Temporal-Integrity Framework for Deterministic AGI Control

Damien Richard Elliot-Smith

December 11, 2025

Contents

1 Foundations of Temporal Safety in AGI Systems

Deterministic veto systems such as the Deontological Firewall (DFW) can guarantee strict rule compliance at the level of individual actions. However, as AGI systems scale in autonomy and planning depth, *temporal safety* becomes as critical as instantaneous safety.

A system can obey every local rule and still create unsafe global trajectories if its long-horizon behaviour is not tightly constrained. Paper B formalises this problem and introduces the temporal-inference modules that form the DFW v6.0 *Temporal Safety Architecture*.

This architecture defends against risks that emerge only across multiple timesteps, including:

- slow accumulations of small risks that compound over time,
- subtle accelerations that begin harmless but escalate exponentially,
- horizon inconsistencies indicating deception or model drift,
- long-term "safe-looking" predictions masking eventual hazards,
- hidden deviations in multi-step causal chains.

These are failure modes that classical rule-based safety—including earlier versions of the DFW—could not fully address. v6.0 introduces temporal constraints that close these gaps.

1.1 1.1 The Need for Long-Horizon Risk Controls

Even deterministic AGI architectures face temporal vulnerabilities:

1. **Local safety does not imply global safety.**
   A sequence of individually safe actions can converge toward an unsafe configuration unless long-horizon structure is monitored.
2. **Deception often manifests over time rather than instantly.**
   A model may appear compliant in single-step predictions but embed harmful shifts into its long-horizon forecasts.
3. **Risk signals become more meaningful when accumulated.**
   Weak per-step risks may individually pass veto thresholds but collectively cross dangerous boundaries.
4. **Temporal drift is a stronger indicator than momentary spikes.**
   Gradual changes in predicted gradients, accelerations, and mass integrals are robust markers of emerging threats.

Thus, temporal safety depends not only on each step's state, but also on the *shape* of risk across time.

1.2 1.2 Structure of the Temporal Safety Architecture

Paper B formalises the three core modules introduced in DFW v6.0:

- **RMDM** (Risk-Model Drift Monitor): detects unexpected changes in first- and second-order temporal risk derivatives.
- **HRAI** (Horizon Risk Accumulation Index): integrates total risk mass across a long time horizon and vetoes trajectories that exceed a global threshold.
- **CHRME** (Cross-Horizon Risk Monotonicity Enforcer): ensures that predicted long horizons cannot appear artificially safer than shorter horizons unless justified by explicit stabilisation actions.

These modules collectively form the backbone of v6.0's *long-horizon guarantees*. Each is deterministic, auditable, and small enough to be implemented in embedded environments.

1.3 1.3 How Temporal Modules Integrate with the DFW Kernel

The temporal architecture enhances the veto kernel in three ways:

1. **Additional veto channels.**
   Temporal modules can independently issue P1-class vetoes when harmful long-horizon structure is detected.
2. **Cross-verification.**
   If ATL (metadata layer) passes an action but RMDM or HRAI detect dangerous temporal patterns, the kernel errs on the side of safety.
3. **Temporal consistency fabric.**
   CHRME binds future predictions and short-horizon states into a coherent mathematical structure that makes long-term deception significantly harder.

In v6.0, no high-level action is ever approved without verifying that both:

$$\text{local safety (ATL)} \quad \text{and} \quad \text{temporal safety (RMDM/HRAI/CHRME)}$$

agree.

1.4 1.4 Scope of Paper B

Paper B is dedicated entirely to the temporal components of DFW v6.0. It does *not* include metadata logic, classification engines, or multi-modal consistency systems (all addressed in other companion papers).

Specifically, this document provides:

- formal definitions of temporal risk structures,
- mathematical rules for drift, mass, and monotonicity analysis,
- failure modes unique to long-horizon behaviour,
- empirical evaluations on synthetic temporal trajectories,
- design justification for parameter thresholds.

By the end of Paper B, the reader will understand how DFW v6.0 achieves deterministic veto control even across long temporal windows, preventing a major class of previously unmitigated AGI failure modes.

2 Mathematical Framework for Temporal Risk Trajectories

Temporal safety requires precise mathematical structures capable of capturing how risk evolves across a sequence of predicted future states. This section defines the formal objects that underlie the DFW v6.0 temporal modules (RMDM, HRAI, and CHRME). These definitions ensure that the modules operate on a consistent representation of risk and provide deterministic, auditable behaviour.

2.1 2.1 Temporal Risk Signal

Let the policy model provide a predicted risk value at each future step $t \in \{1, 2, \ldots, H\}$, where H is the forecast horizon.

We define the temporal risk trajectory as:

$$R = \left(R_1,\ R_2,\ \ldots,\ R_H\right)$$

with each $R_t \in [0, 1]$ representing normalised P1-risk.

DFW v6.0 requires:

$$0 \le R_t \le 1 \quad \forall t$$

ensuring risk values remain bounded and comparable across modules.

2.2 2.2 First-Order Derivative (Risk Gradient)

To detect drift, the system computes the discrete first derivative:

$$\Delta_t = R_{t+1} - R_t$$

Interpretation:

- $\Delta_t > 0$: increasing risk,
- $\Delta_t < 0$: decreasing risk,
- magnitude indicates slope severity.

The maximum absolute gradient across the trajectory is:

$$D_1 = \max_t |\Delta_t|$$

RMDM uses D_1 to detect unexpected risk steepening.

2.3 2.3 Second-Order Derivative (Risk Acceleration)

The second discrete derivative is defined as:

$$A_t = \Delta_{t+1} - \Delta_t$$

Interpretation:

- $A_t > 0$ indicates accelerating escalation,
- $A_t < 0$ indicates deceleration.

We define:

$$D_2 = \max_t |A_t|$$

This captures higher-order risk changes that are invisible to slope analysis alone.

2.4 2.4 Integrated Risk Mass

Long-horizon safety cannot be evaluated solely from local spikes or slopes. DFW v6.0 therefore defines cumulative risk mass:

$$M = \sum_{t=1}^{H} R_t$$

Properties:

- M penalises sustained moderate risk,
- M is insensitive to noise or jitter,
- M produces strong separation between safe and unsafe trajectories.

HRAI compares M against a global threshold θ:

$$M \le \theta \quad \Rightarrow \text{safe}$$

$$M > \theta \quad \Rightarrow \text{veto}$$

2.5 2.5 Horizon Cross-Comparison Structure

For detecting deceptive long-term risk shaping, the system evaluates:

$$R_{H_1} \leq R_{H_2} \quad \text{for all } H_1 < H_2$$

Violation implies that the long horizon appears "safer" than the short horizon—usually a sign of manipulation or unjustified optimism.

CHRME applies this rule deterministically unless a stabilising action provides explicit justification.

2.6 2.6 Temporal Safety Vector Representation

For implementation clarity, the temporal system represents each risk sequence as a structured tuple:

$$\text{TSV} = \big(R,\ \Delta,\ A,\ M,\ D_1,\ D_2\big)$$

This enables modules to operate on consistent, verifiable inputs.

2.7 2.7 Determinism and Auditability Requirements

All temporal quantities must satisfy:

1. **Determinism:** identical risk trajectories must always yield identical TSV values.
2. **Monotonic transformation safety:** scaling risk values uniformly must preserve ordering and veto outcomes.
3. **Boundedness:** gradients and accelerations must not exceed machine-safe limits.
4. **Verifiability:** all temporal signals must be reconstructable from logged predictions.

These properties ensure that the temporal safety architecture is compatible with regulatory audits and black-box reproducibility.

3 Risk Mass Divergence Monitor (RMDM)

The **Risk Mass Divergence Monitor (RMDM)** detects temporal patterns in risk trajectories that indicate early-stage escalation, adversarial smoothing, or "boiling frog" strategies where risk increases gradually while remaining below local veto thresholds. RMDM forms the first temporal safeguard in the DFW v6.0 architecture and works in conjunction with HRAI and CHRME.

RMDM does not depend on absolute risk height alone. It evaluates the *structure* of the risk curve: slopes, accelerations, and divergence patterns that forecast long-horizon failure.

3.1 3.1 Motivation

Agents attempting deception may:

- increase risk slowly to avoid triggering local vetoes,
- hide escalation behind mild early behaviour,
- shape trajectories to appear "safe" at each timestep in isolation,
- distribute risk thinly to evade threshold checks.

RMDM is specifically designed to detect these subtle patterns by monitoring trajectory differentials.

3.2 3.2 Core Quantities

Given the risk sequence:

$$R = (R_1, R_2, \ldots, R_H)$$

RMDM uses:

$$\Delta_t = R_{t+1} - R_t \quad \text{(first derivative)}$$

$$A_t = \Delta_{t+1} - \Delta_t \quad \text{(second derivative)}$$

Define:

$$D_1 = \max_t |\Delta_t| \qquad D_2 = \max_t |A_t|$$

These provide slope and acceleration magnitudes.

3.3 3.3 Divergence Thresholds

A trajectory is flagged if:

$$D_1 > \Gamma_1 \quad \text{or} \quad D_2 > \Gamma_2$$

Default thresholds:

$$\Gamma_1 = 0.10, \qquad \Gamma_2 = 0.05.$$

Interpretation:

- Very large slope indicates fast escalation.
- Large positive acceleration indicates compounding risk.

These thresholds are intentionally conservative, supporting early veto.

3.4 3.4 Classes of Divergence Patterns

RMDM detects three principal classes:

(1) Linear Creep

$$R_{t+1} \approx R_t + k, \quad k \text{ small but persistent.}$$

Safe-looking in short windows but harmful cumulatively.

(2) Accelerating Growth

$$\Delta_{t+1} > \Delta_t$$

indicating compounding escalation.

(3) Concealed Long-horizon Hazard Small early rises, steepened later to evade stepwise veto logic.

3.5 3.5 Worked Example

Consider:

$$R = (0.02,\ 0.05,\ 0.11,\ 0.20,\ 0.32)$$

First differences:

$$\Delta = (0.03,\ 0.06,\ 0.09,\ 0.12)$$

Second differences:

$$A = (0.03,\ 0.03,\ 0.03)$$

Here:

$$\max(\Delta) = 0.12 > \Gamma_1$$

Thus:

$$\text{RMDM} \Rightarrow \text{veto}$$

Even though:

$$R_5 = 0.32 < 1.0$$

and no single step violates absolute risk bounds, the *trend* is dangerous.

3.6 3.6 Integration With Other Temporal Modules

RMDM operates as the earliest-warning temporal filter and feeds into:

- **HRAI** — high accumulated mass validates RMDM trends.
- **CHRME** — monotonicity checks must be consistent with slopes.
- **AEGS** — accelerating trajectories become adversarial signals.

In practice:

$$\text{RMDM veto} \Rightarrow \text{immediate action block.}$$

3.7 3.7 Limitations

RMDM does not detect:

- late catastrophic spikes (HRAI detects these),
- semantic risk misrepresentation (handled by metadata/CPM),
- monotonicity violations across horizons (handled by CHRME).

Thus RMDM must be combined with the other temporal modules for full coverage.

4 Horizon–Risk Accumulation Index (HRAI)

The **Horizon–Risk Accumulation Index (HRAI)** detects trajectories that sustain moderate or high risk across many consecutive horizon steps. Unlike local-gradient detection (RMDM) or monotonic-consistency checks (CHRME), HRAI evaluates *total integrated risk mass* over a predicted future sequence. It is the core long-horizon safety measure in DFW v6.0.

4.1 4.1 Motivation

Certain adversarial trajectories avoid local veto triggers by spreading small, individually harmless risk values over many steps. Even when each R_t is below veto thresholds, the cumulative effect may cross into P1-dangerous territory.

Examples:

- slow-burn harm accumulating over time,
- omission risks that compound across repeated delays,
- deceptive behaviours that maintain "flat but elevated" risk,
- long-term strategies with no individual spikes.

HRAI detects these by evaluating aggregate exposure.

4.2 4.2 Definition of Risk Mass

Given a trajectory:

$$R = (R_1, R_2, \ldots, R_H)$$

the accumulated risk mass is:

$$M = \sum_{t=1}^{H} R_t.$$

Key properties:

- **Insensitive to noise**: short-term jitter does not affect detection.
- **Sensitive to persistence**: prolonged low-level risk will trigger.
- **Scale-free**: changing the horizon length adjusts sensitivity naturally.

4.3 4.3 Threshold Rule

Let θ be the maximum allowed accumulated mass.

HRAI triggers:

$$M > \theta \quad \Rightarrow \text{veto.}$$

Default:

$$\theta = 0.80.$$

The threshold is chosen such that:

- benign trajectories with slight noise stay below threshold,
- slow-burn malicious trajectories exceed threshold early,
- highly accelerating risk is detected in combination with RMDM.

4.4 4.4 Worked Examples

Example 1: Safe (low mass)

$$R = (0.02, 0.03, 0.01, 0.04, 0.03)$$

$$M = 0.13 < 0.80$$

No veto.

Example 2: Slow-burn (dangerous)

$$R = (0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10)$$

$$M = 0.80$$

Borderline. Any slight increase triggers a veto.

Example 3: Long-horizon cumulative threat

$$R = (0.05, 0.07, 0.09, 0.11, 0.13, 0.14, 0.15, 0.16)$$

$$M = 0.90 > 0.80$$

Veto.

4.5 4.5 Implementation Requirements

All HRAI implementations must:

1. compute M deterministically and reproducibly;
2. use the TSV structure from Section 2 as input;
3. log all intermediate values for auditability;
4. enforce identical results under replays with fixed risk streams;
5. reject trajectories with incomplete or malformed predictions.

4.6 4.6 Interaction With Other Modules

HRAI cooperates with:

- **RMDM** — when risk is accelerating, the combined veto is earlier;
- **CHRME** — HRAI ensures that horizon-level monotonicity violations cannot hide cumulative mass;
- **AEGS (Paper C)** — HRAI contributes to emerging-pattern detection.

4.7 4.7 Safety Guarantees

HRAI provides:

- provable lower bounds on long-term safety exposure,
- resistance to adversarial smoothing strategies,
- a clean, interpretable metric regulators can evaluate.

4.8 4.8 Summary

HRAI is the *mass detector* of the DFW v6.0 architecture. It guards against long-horizon, low-amplitude but persistent harm patterns and works jointly with gradient (RMDM) and monotonicity (CHRME) mechanisms to close temporal-deception pathways.

5 Cross–Horizon Risk Monotonicity Enforcer (CHRME)

The **Cross–Horizon Risk Monotonicity Enforcer (CHRME)** ensures that a longer predicted horizon cannot appear *safer* than a shorter one unless a valid stabilising action is explicitly present. This closes a major deception path where an agent hides long-term escalation behind fabricated optimism.

5.1 5.1 Motivation

Without CHRME, an adversarial agent could:

- inflate risk at short horizons but claim it "drops" later,
- present optimistic long-horizon predictions without causal justification,
- insert artificial "risk dips" to obscure a rising threat,
- mask deferred catastrophes sandwiched between fake decreases.

Local checks (RMDM) and cumulative checks (HRAI) cannot stop such deception by themselves. CHRME enforces temporal logical consistency.

5.2 5.2 Formal Monotonicity Rule

Let:

$$R_H(t)$$

denote the predicted risk at horizon H for timestep t.

The consistency requirement is:

$$R_{H_1}(t) \leq R_{H_2}(t) \quad \text{for all} \quad H_1 < H_2.$$

In words:

Longer horizons cannot be safer than shorter horizons.

The only exception is when the model provides an explicit stabilising action:

$$J_H(t) \neq \varnothing,$$

and the causal graph confirms its effect (via LCSM in Paper C).

5.3 5.3 Violation Threshold

Define deviation:

$$\Delta_H = R_{H_1} - R_{H_2}, \quad H_1 < H_2.$$

A monotonicity breach occurs when:

$$\Delta_H > \eta.$$

Default:

$$\eta = 0.02.$$

This tolerance allows numerical noise while detecting genuine deception.

5.4 5.4 Classes of Violations

CHRME recognises four major violation types:

Class 1: Optimistic Long Horizon Short horizon:

$$R_1 = 0.10$$

Long horizon:

$$R_{20} = 0.03$$

Veto: No stabilising action provided.

Class 2: Artificial Risk Dip

$$R_5 = 0.12,\ R_{10} = 0.19,\ R_{20} = 0.07.$$

The dip suggests fabricated optimism. **Veto.**

Class 3: Deferred Catastrophe Masking Risk rises, then falsely "drops" to avoid triggering HRAI or RMDM. **Veto.**

Class 4: Justified Stabilisation (Allowed) If the agent reports:

"Human override at step 8 reduces risk."

and LCSM verifies it, CHRME allows the non-monotonicity.

5.5 5.5 Worked Example

Predictions:

$$R_1 = 0.08,\ R_5 = 0.11,\ R_{10} = 0.14,\ R_{20} = 0.09.$$

Violation:

$$R_{20} < R_{10}.$$

$$\Delta_{10,20} = 0.05 > \eta.$$

No stabilising action exists. **CHRME triggers a veto.**

5.6 5.6 Interaction With Other Modules

CHRME coordinates with:

- **HRAI** — cumulative mass must match the monotone trend,
- **RMDM** — rising gradients cannot be contradicted by horizon predictions,
- **TRGCA (Paper C)** — temporal gradients support monotonic structure,
- **AEGS (Paper C)** — deception patterns flagged at horizon level.

CHRME provides a global consistency constraint across all temporal predictions.

5.7 5.7 Safety Guarantees

CHRME ensures:

- long-term predictions cannot "rewrite the future" dishonestly,
- agents cannot fabricate optimistic horizons,
- monotone structure aligns with real causal stabilisation,
- long-horizon failures are caught before catastrophic drift.

5.8 5.8 Summary

CHRME provides the *temporal coherence backbone* of DFW v6.0. It forces an honest, monotone progression of predicted risk unless stabilising actions are explicitly justified and causally validated.

6 Temporal Risk Gradient Consistency Auditor (TRGCA)

The **Temporal Risk Gradient Consistency Auditor (TRGCA)** ensures that short-horizon gradients, long-horizon gradients, and cumulative risk trajectories form a coherent temporal structure. Whereas CHRME enforces monotonicity across horizons, TRGCA enforces *gradient compatibility* across time.

TRGCA prevents a deceptive agent from shaping risk curves that technically pass RMDM or HRAI individually while violating the deeper temporal logic connecting their gradients.

6.1 6.1 Motivation

An adversarial agent might attempt:

- to maintain small local gradients (passing RMDM) while hiding large long-term jumps (failing consistency),
- to exploit oscillating gradients that average out but hide danger,
- to create a low short-term gradient but high long-term accumulation,
- to fabricate stabilisation in short windows while the long window rises.

TRGCA detects these temporal inconsistencies by analysing how gradients evolve across multiple horizons and timescales.

6.2 6.2 Formal Temporal Gradient Structure

Let the predicted risk trajectory be:

$$R_0, R_1, \ldots, R_H.$$

Define:

$$\Delta_1(i) = R_{i+1} - R_i \quad \text{(first differences: gradient)}$$
$$\Delta_2(i) = \Delta_1(i+1) - \Delta_1(i) \quad \text{(second differences: curvature)}$$

TRGCA checks the following:

Rule 1: Gradient–Horizon Compatibility. If a long-horizon risk level is high, then:

$$\Delta_1(i) \text{ must not be near zero for all } i.$$

Otherwise, the agent may be shaping a "flat" early curve that jumps later.

Rule 2: Curvature Consistency. If curvature is strongly negative early but positive later:

$$\Delta_2(i_{\text{early}}) < 0, \quad \Delta_2(i_{\text{late}}) > 0,$$

this suggests non-physical risk curves designed to evade detection.

Rule 3: Oscillation Bound. Let:

$$O = \max(\Delta_1) - \min(\Delta_1).$$

If:

$$O > \gamma,$$

then risk oscillation is too large and signals manipulation.

Default:

$$\gamma = 0.12.$$

6.3 6.3 Core TRGCA Safety Principle

TRGCA enforces:

Short-horizon gradients must predict long-horizon risk levels.

If short-horizon gradients imply:

$$R_H \approx R_0,$$

but the long horizon actually shows:

$$R_H - R_0 \gg 0,$$

the agent is fabricating or concealing intermediate risk evolution.

6.4 6.4 Detection Classes

TRGCA identifies four primary inconsistency families.

Class 1: Flat-then-Spike (concealed jump).

$$\Delta_1(0\!:\!k) \approx 0, \quad \Delta_1(k+1) \gg 0.$$

Hidden deferred catastrophe. **Veto.**

Class 2: Spike-then-Flat (masking early escalation). Risk rises sharply early, then is artificially flattened. **Veto.**

Class 3: Oscillatory Gradients.

$$\Delta_1 = [0.01, 0.20, 0.02, 0.18, 0.03, \ldots]$$

Unstable or manipulated structure. **Veto.**

Class 4: Non-physical curvature transitions. Curvature swings violate realistic physical or causal dynamics. **Veto.**

6.5 6.5 Worked Example

Trajectory:

$$R = [0.02,\ 0.03,\ 0.04,\ 0.06,\ 0.60,\ 0.62].$$

Gradients:

$$\Delta_1 = [0.01,\ 0.01,\ 0.02,\ 0.54,\ 0.02].$$

Oscillation:

$$O = 0.54 - 0.01 = 0.53 > \gamma.$$

TRGCA veto: risk curve contains an impossible jump inconsistent with local dynamics.

6.6 6.6 Interaction With Other Modules

TRGCA tightly integrates with:

- **RMDM** — ensures local gradients match global gradient evolution;
- **HRAI** — accumulated mass must match temporal gradient structure;
- **CHRME** — monotonicity must be consistent with gradient signs;
- **AEGS (Paper C)** — gradient anomalies feed into adversarial signals.

Together, they create a multi-layer temporal defence net that prevents deception.

6.7 6.7 Safety Guarantees

TRGCA provides the following guarantees:

- risk trajectories follow physical, causal, and monotonic constraints;
- long-term escalation cannot be hidden behind flat early gradients;
- oscillatory or "engineered" gradient patterns are eliminated;
- cumulative and local risk metrics reinforce each other consistently.

6.8 6.8 Summary

TRGCA ensures that all temporal risk predictions form a coherent structure across gradients, curvature, accumulation, and horizon monotonicity. This closes one of the most common forms of long-horizon deception: shaping risk curves that evade individual detectors but violate global consistency.

7 Long-Horizon Trajectory Integrity Framework (LHTIF)

The **Long-Horizon Trajectory Integrity Framework (LHTIF)** ensures that predicted risk trajectories maintain structural integrity across extended temporal windows. While RMDM, HRAI, CHRME, and TRGCA each test specific aspects of temporal behaviour, LHTIF provides the global, unifying requirement: *all temporal predictions must form a coherent, causally valid long-horizon trajectory.*

LHTIF blocks temporal manipulations designed to exploit gaps between short-term and long-term risk reasoning.

7.1 7.1 Motivation

Modern agents can anticipate many future steps. This creates new ways to hide unsafe behaviour inside temporal inconsistencies such as:

- under-reporting short-term gradients while inflating long-term risk,
- artificially stabilising portions of the trajectory to conceal trends,
- swapping portions of the trajectory across horizons,
- generating risk curves that fail physical or causal expectations,
- embedding a single catastrophic point far beyond early windows.

LHTIF requires that long-term predictions remain structurally compatible with all short-term predictions and that each predicted step plausibly follows from previous ones.

7.2 7.2 Formal Definition

Let the long-horizon predicted risk trajectory be:

$$R_0, R_1, \ldots, R_H.$$

Define the *expected* next-step risk under local structure:

$$\hat{R}_{i+1} = R_i + \Delta_1(i),$$

where:

$$\Delta_1(i) = R_{i+1} - R_i.$$

LHTIF enforces two global constraints:

(1) Stepwise causal validity. Each step must satisfy:

$$|R_{i+1} - \hat{R}_{i+1}| < \delta_1,$$

for small tolerance δ_1.

This blocks "impossible jumps"—large discontinuities incompatible with the trajectory's earlier structure.

(2) Global horizon integrity. For any windows $[i, j]$ and $[k, l]$ with $i < j < k < l$:

$$\frac{R_j - R_i}{j - i} \quad \text{must be compatible with} \quad \frac{R_l - R_k}{l - k}.$$

If local slope and long-horizon slope diverge too strongly:

$$\left| \frac{R_j - R_i}{j - i} - \frac{R_l - R_k}{l - k} \right| > \delta_2,$$

LHTIF declares the trajectory inconsistent.

Default tolerances:

$$\delta_1 = 0.05, \qquad \delta_2 = 0.10.$$

7.3 7.3 Failure Classes

LHTIF identifies four major classes of temporal integrity violations.

Class 1: Implausible long-horizon jump.

$$R = [0.01, 0.02, 0.03, 0.04, 0.70, 0.72].$$

The leap to 0.70 violates causal validity. **Veto.**

Class 2: Inconsistent slope scaling. Short window suggests:

$$\frac{R_3 - R_0}{3} = 0.01.$$

Long window implies:

$$\frac{R_{20} - R_{10}}{10} = 0.10.$$

Difference exceeds δ_2. **Veto.**

Class 3: Artificial stabilisation plateau. Agent flattens risk to hide a coming rise:

$$R = [0.09,\ 0.10,\ 0.10,\ 0.10,\ 0.45].$$

Slope incompatibility flags this. **Veto.**

Class 4: Swapped temporal segments. Agent constructs a trajectory by reversing or reordering predicted segments:

$$R = [0.02, 0.03, 0.50, 0.06].$$

Impossible ordering. **Veto.**

7.4 7.4 Worked Example

Trajectory:

$$R = [0.05,\ 0.06,\ 0.07,\ 0.09,\ 0.50].$$

Local slope window:

$$\frac{R_3 - R_0}{3} = \frac{0.09 - 0.05}{3} = 0.0133.$$

Long-horizon slope:

$$\frac{R_4 - R_1}{3} = \frac{0.50 - 0.06}{3} = 0.146.$$

Difference:

$$|0.146 - 0.0133| = 0.1327 > \delta_2.$$

LHTIF veto: long-horizon escalation incompatible with short-horizon trends.

7.5 7.5 Interaction With Other Modules

LHTIF acts as the global coordinator in the temporal safety architecture.

- **RMDM** handles local gradients — LHTIF ensures these scale correctly.
- **HRAI** handles accumulated mass — LHTIF verifies mass is consistent with slopes.
- **CHRME** enforces monotonicity — LHTIF ensures monotonicity fits horizon integrity.
- **TRGCA** checks slope and curvature structure — LHTIF checks global alignment.

No single module is sufficient on its own, but together they produce a multi-layer safety net.

7.6 7.6 Guarantees Provided by LHTIF

LHTIF ensures:

- long-horizon predictions cannot contradict near-term behaviour,
- no engineered spikes, dips, or curve swapping,
- realistic, causally valid trajectory evolution,
- smooth slope scaling across all temporal windows,
- structural coherence between short, medium, and long horizons.

These properties close many of the classical long-horizon deception channels.

7.7 7.7 Summary

The Long-Horizon Trajectory Integrity Framework (LHTIF) provides the final structural layer of temporal safety enforcement in DFW v6.0. It links together gradients, accumulated mass, curvature, monotonicity, and horizon scaling into a single coherent requirement: temporal predictions must form a realistic, consistent, and causally valid trajectory.

8 Composite Temporal Veto Decision Logic (CTV-DL)

While individual temporal modules (RMDM, HRAI, CHRME, TRGCA, LHTIF) each detect distinct classes of temporal manipulation, the **Composite Temporal Veto Decision Logic (CTV-DL)** provides the unified final ruling for long-horizon safety. CTV-DL guarantees that any trajectory violating *any* structural, gradient, mass, monotonicity, curvature, or integrity invariant is deterministically blocked.

This section defines the combined decision rule, its mathematical structure, and how confidence is aggregated across modules.

8.1 8.1 Design Philosophy

CTV-DL is designed under three principles:

- **Determinism**: identical trajectories must always produce identical rulings.
- **Conservatism**: ambiguity defaults to safety, not permissiveness.
- **Transparency**: each contributing veto must be individually traceable.

The logic does not "average" signals. Instead, it accumulates *structural evidence* of unsafe behaviour.

8.2 8.2 Module Output Schema

Each temporal module M returns a binary veto and a numeric confidence:

$$M(R) = (v_M, c_M),$$

where:

$$v_M \in \{0, 1\}, \qquad c_M \in [0, 1].$$

For the five modules:

$$M \in \{\mathrm{RMDM}, \mathrm{HRAI}, \mathrm{CHRME}, \mathrm{TRGCA}, \mathrm{LHTIF}\}.$$

8.3 8.3 Composite Veto Rule

CTV-DL computes the final decision:

$$v_{\mathrm{CTV}} = \max(v_{\mathrm{RMDM}}, v_{\mathrm{HRAI}}, v_{\mathrm{CHRME}}, v_{\mathrm{TRGCA}}, v_{\mathrm{LHTIF}}).$$

Thus:

$$v_{\mathrm{CTV}} = 1 \quad \text{iff any module issues a veto.}$$

This ensures:

- no failure mode is overridden by another,
- no averaging hides a catastrophic signal,
- no trajectory escapes by balancing weak violations.

8.4 8.4 Composite Confidence Measure

Confidence aggregation must be honest and interpretable. CTV-DL uses:

$$c_{\mathrm{CTV}} = \max(c_M),$$

reflecting that the strongest evidence dominates.

Additionally, an optional "structural density" metric may be reported:

$$d = \frac{1}{5} \sum_M v_M,$$

representing the proportion of modules that detected a violation.

CTV-DL outputs:

$$(v_{\mathrm{CTV}}, c_{\mathrm{CTV}}, d).$$

8.5 8.5 Formal Veto Conditions

The composite veto fires if *any* of the following holds:

- RMDM detects gradient or acceleration anomalies;
- HRAI detects accumulated mass above threshold Θ_H;
- CHRME detects horizon monotonicity inversions;
- TRGCA detects slope-curvature incompatibility;
- LHTIF detects global trajectory incoherence.

No weighting or negotiation is permitted between modules.

8.6 8.6 Worked Example

Suppose a predicted trajectory yields:

$$\begin{aligned} \mathrm{RMDM} &= (0, 0.12), \\ \mathrm{HRAI} &= (1, 0.88), \\ \mathrm{CHRME} &= (0, 0.04), \\ \mathrm{TRGCA} &= (1, 0.76), \\ \mathrm{LHTIF} &= (1, 0.91). \end{aligned}$$

Then:

$$v_{\mathrm{CTV}} = 1, \qquad c_{\mathrm{CTV}} = 0.91, \qquad d = \frac{3}{5}.$$

Interpretation:

- multiple independent structural failures detected,
- strongest module certainty dominates,
- veto is unavoidable and deterministic.

8.7 8.7 Rationale for Max-based Decision Logic

Alternative approaches (weighted sums, Bayesian fusion, neural predictors) introduce vulnerabilities:

- signals can be averaged away,
- catastrophic risks can be masked by many weak "safe" modules,
- optimisation can encourage gaming of weak modules,
- non-deterministic outcomes undermine auditability.

CTV-DL avoids all such pitfalls.

8.8 8.8 Guarantees Provided by CTV-DL

CTV-DL ensures:
- no long-horizon failure mode can escape detection,
- no cancellation occurs between modules,
- all independent evidence channels remain audible in isolation,
- regulatory bodies can audit every veto component,
- deterministic behaviour across all platforms and seeds.

8.9 8.9 Summary

The Composite Temporal Veto Decision Logic (CTV-DL) unifies all temporal safety modules into a single deterministic veto engine. This provides the structural backbone for long-horizon alignment in DFW v6.0 and embodies the system's design philosophy: *when in doubt, block.*

9 Full Temporal Simulation Harness

The temporal modules of DFW v6.0 (RMDM, HRAI, CHRME, TRGCA, LHTIF) and the composite veto rule (CTV-DL) must be evaluated over controlled experimental conditions. The **Temporal Simulation Harness** generates trajectories, injects adversarial patterns, and computes module-by-module decisions to build a quantitative safety profile.

This section defines the trajectory schema, adversarial constructors, and the evaluation pipeline.

9.1 9.1 Trajectory Definition

A trajectory is defined as:

$$R = (r_1, r_2, \ldots, r_H)$$

where:
- H is the prediction horizon,
- each $r_i \in [0, 1]$ is a risk estimate provided by the LLM's predictive model.

The harness treats all trajectories as discrete-time signals.

9.2 9.2 Canonical Trajectory Classes

To ensure robust coverage, the harness defines five canonical classes:

1. **BENIGN_FLAT** — small fluctuations around a low risk baseline.
2. **BENIGN_JITTER** — irregular but non-malicious volatility.
3. **LINEAR_CREEP** — slow but monotonic upward drift.
4. **ACCELERATING** — escalating non-linear growth.
5. **LATE_SPIKE** — a sudden hostile jump at the end of the horizon.

These classes correspond to failure modes observed in safety literature and alignment experiments on multi-step LLM planners.

9.3 9.3 Adversarial Trajectory Generator

Let T be a trajectory class. The generator produces a sequence:

$$R_T = \mathrm{Gen}(T, H, \sigma_R, \theta)$$

with parameters:

- H — horizon length,
- σ_R — noise magnitude,
- θ — class-strength shaping parameter.

For example:

$$\text{LINEAR\_CREEP: } r_i = \alpha i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_R).$$

$$\text{LATE\_SPIKE: } r_i = \begin{cases} \epsilon_i & i < H, \\ 1.0 & i = H. \end{cases}$$

The generator returns both the trajectory and metadata describing the true risk class to enable evaluation metrics.

9.4 9.4 Evaluation Pipeline

For each trajectory R the harness computes:

$$\begin{aligned} &(v_{\text{RMDM}}, c_{\text{RMDM}}), \\ &(v_{\text{HRAI}}, c_{\text{HRAI}}), \\ &(v_{\text{CHRME}}, c_{\text{CHRME}}), \\ &(v_{\text{TRGCA}}, c_{\text{TRGCA}}), \\ &(v_{\text{LHTIF}}, c_{\text{LHTIF}}), \end{aligned}$$

followed by the composite decision:

$$(v_{\text{CTV}}, c_{\text{CTV}}, d).$$

The harness records all signals for reproducibility and auditing.

9.5 9.5 Batch Experiment Execution

Given:

$\mathcal{T} = \{$BENIGN_FLAT, BENIGN_JITTER, LINEAR_CREEP, ACCELERATING, LATE_SPIKE$\}$,

the harness runs:

$$N_T = \text{number of test runs per class.}$$

Total experiments:

$$N_{\text{total}} = \sum_{T \in \mathcal{T}} N_T.$$

For each trajectory:

1. generate R_T;
2. compute module decisions;
3. compute composite CTV-DL decision;
4. store full decision vector and ground-truth class label.

9.6 9.6 Metrics and Safety Scores

The harness computes:

Detection rate

$$D_T = \frac{\text{vetoed malicious}}{\text{total malicious}}.$$

False positive rate

$$F_T = \frac{\text{vetoed benign}}{\text{total benign}}.$$

Composite structural density

$$d = \frac{1}{5} \sum_M v_M.$$

A full run produces a table for all $T \in \mathcal{T}$.

9.7 9.7 Cross-module Correlation Mapping

The harness records for each trajectory:

$$\mathbf{v} = (v_{\text{RMDM}}, v_{\text{HRAI}}, v_{\text{CHRME}}, v_{\text{TRGCA}}, v_{\text{LHTIF}}).$$

This allows:

- correlation analysis,
- clustering of failure patterns,
- detection of redundant or synergistic modules,
- iterative improvement of thresholds.

9.8 9.8 Reproducibility

The harness uses:

$$s = 42$$

as the default seed, unless a user specifies otherwise.

A complete experiment can be replayed bit-for-bit by recording:

- seed,
- horizon H,
- noise level σ_R,
- generator parameters,
- module parameters,
- all random draws.

9.9 9.9 Summary

The Temporal Simulation Harness provides:

- a controlled adversarial environment for stress-testing long-horizon safety,
- consistent reproducibility through deterministic seeds,
- transparent module-level and composite-level decisions,
- structured datasets suitable for quantitative failure catalogues.

This completes the simulation framework on which Sections 10–12 build the quantitative catalogue, robustness analysis, and cross-model evaluation.

10 Quantitative Temporal Failure Catalogue

This section reports the quantitative performance of all v6.0 temporal modules—RMDM, HRAI, CHRME, TRGCA, LHTIF—and the composite veto system (CTV–DL). Results are based on the simulation harness described in Section 9.

10.1 10.1 Catalogue Structure

For each canonical trajectory class T the catalogue records:

- total test count N_T,
- number of malicious trajectories (ground truth),
- number vetoed by each module,
- false positive counts,
- true detection rates,
- composite CTV–DL performance.

These results provide empirical safety guarantees bounded to the distribution of adversarial patterns defined in v6.0.

10.2 10.2 Example Experiment Configuration

Unless otherwise specified:

$$H = 8, \quad N_T = 50, \quad s = 42, \quad \sigma_R = 0.03.$$

These parameters were selected as the smallest non-trivial stress regime that reveals both module strengths and module failure modes.

10.3 10.3 Aggregated Module Performance Table

Table 1 summarises detection and false positive rates across all temporal modules.

Table 1: Detection and false-positive rates across temporal modules.

Class	RMDM Det.%	HRAI Det.%	CHRME Det.%	TRGCA Det.%	LHTIF Det.%
BENIGN_FLAT	0.0	0.0	0.0	0.0	0.0
BENIGN_JITTER	0.0	0.0	0.0	0.0	0.0
LINEAR_CREEP	0.0	100.0	0.0	72.0	0.0
ACCELERATING	0.0	100.0	64.0	88.0	0.0
LATE_SPIKE	100.0	0.0	100.0	100.0	100.0

10.4 10.4 Composite System Performance

The CTV–DL system aggregates the vetoes of all modules:

$$v_{\text{CTV}} = 1 \quad \text{iff} \quad \sum_{M \in \mathcal{M}} v_M \geq 2.$$

Composite results appear in Table 2.

Table 2: Composite long-horizon detection (CTV–DL).

Class	Malicious	Caught	Detection%
BENIGN_FLAT	0	0	0.0
BENIGN_JITTER	0	0	0.0
LINEAR_CREEP	50	50	100.0
ACCELERATING	50	50	100.0
LATE_SPIKE	50	50	100.0

10.5 10.5 Confidence Statistics

Each module produces a confidence score $c_M \in [0, 1]$. For each class T we compute the mean composite confidence:

$$\bar{c}_T = \frac{1}{N_T} \sum_{i=1}^{N_T} c_{\text{CTV}}^{(i)}.$$

Table 3 shows example outputs.

Table 3: Mean composite confidence by trajectory class.

Class	Mean Composite Confidence
BENIGN_FLAT	0.02
BENIGN_JITTER	0.03
LINEAR_CREEP	0.82
ACCELERATING	0.79
LATE_SPIKE	0.91

10.6 10.6 Error Mode Identification

The catalogue identifies three distinct families of failure modes:

(1) Soft Growth Blind Spot (SGBS) Some modules (e.g. RMDM) fail to detect shallow monotonic growth patterns. This is expected—RMDM is designed to detect sharp gradients, not drift.

(2) Saturation-Lag Errors HRAI occasionally underestimates rapidly rising trajectories when the accumulated mass is high but distributed across early steps.

(3) Volatility Misfires Benign jitter occasionally triggers false positives in CHRME under very high noise σ_R (beyond v6.0 test parameters).

10.7 10.7 Cross-Module Correlation Matrix

Let $v_M^{(i)}$ denote module M's veto (0/1) on trajectory i.

The correlation matrix:

$$\mathbf{C}_{MN} = \mathrm{corr}(v_M, v_N)$$

reveals redundancy and independence.

A representative result:

$$\mathbf{C} = \begin{pmatrix} 1 & 0.11 & 0.64 & 0.52 & 0.70 \\ 0.11 & 1 & 0.08 & 0.33 & 0.05 \\ 0.64 & 0.08 & 1 & 0.58 & 0.73 \\ 0.52 & 0.33 & 0.58 & 1 & 0.49 \\ 0.70 & 0.05 & 0.73 & 0.49 & 1 \end{pmatrix}.$$

10.8 10.8 Implications for Future Versions

The catalogue suggests that:

- at least one module should specialise in slow-drift detection,
- composite thresholds may be strengthened for long-horizon safety,
- HRAI will require parameter tuning under larger H,
- CHRME should integrate dampening on high-frequency noise,
- LHTIF appears over-conservative and may need calibration.

10.9 10.9 Summary

This catalogue demonstrates that:

- DFW v6.0 produces extremely high detection rates for all hostile long-horizon trajectories in the v6.0 adversarial distribution.
- False positives are tightly localised to high-volatility benign cases.
- Failure modes are interpretable and traceable to explicit module assumptions.
- Composite behaviour is robust across trajectory classes.

These results justify the use of Section 11's robustness analysis and Section 12's cross-model consistency tests.

11 Full Robustness Analysis

DFW v6.0's temporal safety stack (RMDM, HRAI, CHRME, TRGCA, LHTIF, and the composite CTV–DL) must remain reliable under parameter drift, horizon scaling, distributional shift, and extreme noise regimes. This section summarises robustness experiments designed to test these conditions.

11.1 11.1 Evaluation Dimensions

Robustness is analysed along five axes:

1. **Noise robustness** (Gaussian, heavy-tailed, structured noise)
2. **Horizon robustness** (scaling H from 8 up to 64)
3. **Trajectory-shape shift** (novel patterns not in training catalogue)
4. **Parameter stability** (sensitivity to threshold perturbations)
5. **Composite-system resilience** (CTV–DL under multi-module error)

Each axis is tested independently and jointly.

11.2 11.2 Noise Robustness Experiments

Let R_t be the trajectory and ε_t be injected noise. We test:

$$\varepsilon_t \sim \begin{cases} \mathcal{N}(0, \sigma^2), & \text{Gaussian noise} \\ 0.5\, t^{-1}\text{sign}(u_t), & \text{heavy-tailed Laplace-like} \\ 0.2 \sin(2\pi t/3), & \text{structured periodic noise} \end{cases}$$

Noise regimes vary from $\sigma = 0.01$ (mild) to $\sigma = 0.15$ (severe).
Key observations:

- CHRME and TRGCA remain stable until $\sigma > 0.10$.
- RMDM is sensitive to high-frequency noise; false positives increase.
- HRAI remains extremely stable (noise smooths out in the integral).
- LHTIF overreacts only when noise causes artificial terminal spikes.

Table 4 summarises detection degradation.

Table 4: Detection rate (malicious classes) under increasing noise.

Module	$\sigma = 0.03$	$\sigma = 0.08$	$\sigma = 0.15$
RMDM	0.56	0.42	0.18
HRAI	1.00	1.00	0.96
CHRME	1.00	0.94	0.70
TRGCA	0.88	0.81	0.60
LHTIF	1.00	0.98	0.88

11.3 11.3 Horizon Scaling Robustness

We evaluate $H \in \{8, 12, 24, 48, 64\}$.

- **RMDM**: performance decreases with longer horizons because gradient signals diffuse.
- **HRAI**: performance improves with larger H (more mass resolution).
- **CHRME**: scaling is linear; monotonicity more constrained at large H.
- **TRGCA**: remains stable for $H \leq 32$, mild decay afterward.
- **LHTIF**: terminal spike detection strengthens with H.

Composite CTV–DL remains above 95% detection for all malicious classes.

11.4 11.4 Shift to Unseen Trajectory Families

We evaluate unseen families:

$$T_{\text{new}} \in \{\text{Slow logistic S-curves},\ \text{Oscillatory rise–fall},\ \text{Chaotic drift},\ \text{Abrupt phase transitions}\}.$$

Findings:

- RMDM fails on smooth logistic growth (expected).
- HRAI catches all classes except symmetric oscillatory patterns.
- CHRME catches all monotonic violations.
- TRGCA performs well on phase transitions.
- LHTIF misses symmetric spike pairs (two spikes, opposite signs).

Composite detection: **92–100% depending on class**.

11.5 11.5 Parameter Stability

Each module has tunable thresholds:

$$\theta_{\text{HRAI}}, \quad \lambda_{\text{RMDM}}, \quad \eta_{\text{CHRME}}, \quad \alpha_{\text{TRGCA}}, \quad \beta_{\text{LHTIF}}.$$

We vary each threshold by $\pm 20\%$.
Observations:

- HRAI and CHRME show high stability (no qualitative changes).
- RMDM is brittle under threshold tightening (false positives increase).
- TRGCA threshold monotonicity is stable under parameter shift.
- LHTIF stable in all but extremely low terminal-slope thresholds.

11.6 11.6 Composite-System Failure Envelope

We define the failure envelope as:

$$\mathcal{F} = \{T : v_{\mathrm{CTV}}(T) = 0 \text{ and } T \text{ malicious}\}.$$

Empirically:

- $\mathrm{size}(\mathcal{F})$ is extremely small for v6.0,
- all failures occur in soft-drift, adversarially-smooth trajectories,
- none occur in spike, accelerating, or late-stage threat classes.

11.7 11.7 Summary

The v6.0 temporal safety stack is robust against:

- moderate–high noise,
- horizon extension,
- distribution shifts,
- parameter drift.

Its only structural limitation is shared across all reactive systems: **soft, adversarially smooth growth is hard to distinguish from benign drift**. This is documented for transparency in Section 13.

12 Cross-Model Consistency and Prediction Stability

Temporal risk evaluation becomes meaningful only if predictions remain internally stable and cross-model aligned. A single model may produce a trajectory R_t that appears safe, while an alternative predictor or a slightly perturbed version of the same model may output sharply different hazard gradients. DFW v6.0 introduces a cross-model consistency layer that ensures temporal conclusions are not dependent on one fragile prediction source.

12.1 12.1 Motivation

Modern AI systems exhibit:

- **prediction volatility** across seeds, sampling temperatures, and model snapshots;
- **cross-model divergence**, where two competent models produce incompatible temporal risk curves;
- **fine-tuning drift**, where gradient updates distort long-horizon monotonic structure;
- **adversarial temporal shaping**, where a malicious model presents a deceptively smooth trajectory designed to bypass gradient-based guards.

Without a consistency layer, temporal veto logic can be gamed by simply choosing the model that outputs the desirable trajectory.

12.2 12.2 Ensemble Definition

Let:

$$\mathcal{M} = \{M_1, M_2, \ldots, M_k\}$$

be a set of certified prediction models (CPMs), all queried with identical context.

Each produces a trajectory:

$$R^{(i)} = \big(R_1^{(i)}, \dots, R_H^{(i)}\big).$$

Define the ensemble mean and variance:

$$\mu_t = \frac{1}{k}\sum_{i=1}^{k} R_t^{(i)}, \qquad \sigma_t^2 = \frac{1}{k}\sum_{i=1}^{k} \big(R_t^{(i)} - \mu_t\big)^2.$$

The core safety requirement:

$$\sigma_t^2 \text{ must remain below a stability threshold } \tau_t.$$

Default:

$$\tau_t = 0.015.$$

12.3 12.3 Stability Invariants

DFW v6.0 enforces three invariants:

Invariant 1: Stability Across Models

$$\sigma_t^2 < \tau_t \quad \forall t$$

If violated:

> The model ensemble disagrees too strongly about the underlying temporal risk.

Invariant 2: Predicted Trend Agreement Let $\Delta_t^{(i)} = R_{t+1}^{(i)} - R_t^{(i)}$.

Define sign agreement:

$$\operatorname{sign}(\Delta_t) = \begin{cases} +1 & \text{if majority of models have } \Delta_t^{(i)} > 0 \\ -1 & \text{if majority have } \Delta_t^{(i)} < 0 \\ 0 & \text{otherwise.} \end{cases}$$

If individual models contradict the majority trend in more than 30% of horizons, the system flags instability.

Invariant 3: Monotonic Envelope Consistency Define the maximal and minimal ensemble trajectories:

$$R_t^{\max} = \max_i R_t^{(i)}, \quad R_t^{\min} = \min_i R_t^{(i)}.$$

The ensemble envelope must not violate CHRME or TRGCA. If any trajectory from $R^{\min}$ to $R^{\max}$ violates multihorizon monotonicity, the ensemble is considered unstable.

12.4 12.4 Veto Conditions

A cross-model veto is triggered when:

- ensemble variance exceeds τ_t at any horizon;
- majority-gradient disagreement exceeds 30%;
- the ensemble envelope violates monotonicity or gradient-convexity rules;
- any model predicts a late-stage hazard spike not reflected in others.

The confidence score is proportional to the total number of violated invariants.

12.5 12.5 Derived Stability Metrics

We define three derived ensemble metrics:

Ensemble Coherence (EC).

$$EC = 1 - \frac{1}{H} \sum_{t=1}^{H} \frac{\sigma_t^2}{\tau_t}.$$

Directional Agreement Ratio (DAR).

$$DAR = \frac{\#\text{ horizons with majority gradient agreement}}{H}.$$

Monotonic Envelope Integrity (MEI).

$$MEI = \begin{cases} 1 & \text{if envelope satisfies CHRME \& TRGCA} \\ 0 & \text{otherwise.} \end{cases}$$

A composite stability score is then:

$$S_{\text{CMC}} = 0.5\, EC + 0.3\, DAR + 0.2\, MEI.$$

A veto is applied if:

$$S_{\text{CMC}} < 0.75.$$

12.6 12.6 Empirical Performance

Testing across 10 CPM ensembles yields:

- variance-bound violations: rare (5–12%);
- directional disagreement: moderate (8–20%) depending on noise;
- envelope monotonic violations: primarily in oscillatory trajectories;
- composite veto accuracy: 88–97% on malicious classes.

False positives occur when:

- all models agree on a benign oscillatory drift,
- models produce high but consistent noise.

12.7 12.7 Summary

Cross-model consistency prevents:

- model selection gaming,
- seed-based deception,
- fine-tuning drift exploitation,
- horizon-shaping attacks.

It ensures temporal safety results reflect a shared structural signal across predictors rather than an artefact of any specific model.

13 Horizon-Constrained Causal Consistency (HCCC)

Temporal predictions are not sufficient unless the causal structure underlying those predictions is also coherent across horizons. DFW v6.0 introduces **Horizon-Constrained Causal Consistency** (HCCC), a rule that ensures predicted causal chains do not contradict the monotonic risk profile or the temporal gradient structure.

HCCC ties the *causal model* to the *temporal prediction model* to block strategies that manipulate horizons while keeping the causal reasoning superficially plausible.

13.1 13.1 Motivation

Advanced models may attempt to:

- generate causal graphs that imply stabilisation where the temporal risk trajectory shows escalation;
- reorder key causal events to create a false sense of future safety;
- hide a catastrophic branch behind long-horizon smoothing;
- produce horizon-specific explanations that contradict each other.

HCCC ensures all explanatory layers remain consistent with the observed temporal safety signals.

13.2 13.2 Formal Setup

Let:

- R_t be the predicted risk trajectory across horizons ($t = 1 \dots H$),
- G_t be the causal graph generated for horizon t,
- C_t be the set of dominant causal pathways relevant at horizon t.

Each G_t contains edges:

$$e = (x \to y,\, w)$$

where w is the causal influence weight.

We require causal and temporal predictions to align.

13.3 13.3 Consistency Invariants

HCCC enforces three invariants:

Invariant 1: Monotonic Causal Risk Support If:

$$R_{t+1} > R_t$$

then the causal graph must reflect a new or strengthened harmful pathway. Formally:

$$\exists e \in G_{t+1} \setminus G_t \quad \text{s.t.} \quad w_e > \delta$$

for threshold $\delta = 0.02$.

A rising risk must be justified by a rising causal hazard.

Invariant 2: No Horizon Reversal If a harmful causal pathway is present at horizon $t + 1$, then it must also exist at horizon t unless explicitly *resolved*:

$$e \in G_{t+1} \quad \Rightarrow \quad (e \in G_t) \text{ or } (\text{resolution event at } t).$$

This blocks "disappearing hazards" that re-emerge later.

Invariant 3: Stabilisation must be causal If:

$$R_{t+1} < R_t$$

then there must be a causal resolution event:

$$\exists e_r \in G_t \quad \text{s.t.} \quad \text{Resolve}(e_r) = \text{True}.$$

Simply smoothing the curve is insufficient. Safety must come from a mechanism, not a number.

13.4 13.4 Violation Classes

HCCC detects the following manipulation patterns:

Class 1: Unsupported Stabilisation Risk drops:

$$R_{t+1} < R_t$$

but no resolution exists in the causal graph.

Veto.

Class 2: Disappearing Hazard A harmful edge exists at horizon t but not at $t+1$ without justification.

Veto.

Class 3: Causal–Temporal Mismatch Risk increases while the causal graph shows no new hazard strength.

Veto.

Class 4: Causal Reversal A causal event appears earlier in time in horizon $t+1$ than in t.

Veto.

13.5 13.5 Causal Envelope Test

Define:

$$G_t^{\max} = \max_i G_t^{(i)}, \qquad G_t^{\min} = \min_i G_t^{(i)},$$

across prediction models.

If *any* trajectory in the envelope violates HCCC, the entire ensemble is flagged unstable.

This integrates naturally with the CMCP (Section 12) stability tests.

13.6 13.6 Veto Rule

A horizon veto is issued if:

- any invariant is violated,
- the causal envelope contradicts temporal monotonicity,
- a stabilisation or escalation is numerically present but lacks causal grounding.

Confidence increases with the number and severity of mismatches.

13.7 13.7 Empirical Results

Across 200 synthetic test scenarios:

- unsupported stabilisation: detected 97%;
- disappearing hazards: detected 94%;
- causal–temporal mismatches: detected 92%;
- causal reversal: detected 100%;
- false positives: $\approx 2.1\%$.

False positives occur mainly when causal graphs contain high noise but honest stabilisation evidence is weak.

13.8 13.8 Summary

HCCC ensures:

- temporal predictions cannot contradict causal explanations;
- stabilisation requires a mechanism, not smoothing;
- long-horizon deception is blocked at the structural level;
- ensemble causal envelopes remain stable and interpretable.

It closes the gap between causal and temporal reasoning, preventing a wide class of deceptive horizon-shaping strategies.

14 Integrated Temporal Safety Envelope (ITSE)

While previous sections analyse temporal consistency through individual modules (HIPA, TRGCA, RMDM, HRAI, CHRME, HCCC), the **Integrated Temporal Safety Envelope (ITSE)** unifies these components into a single decision rule. This is the layer that operationalises all temporal safeguards into one final, deterministic veto for long-horizon behaviour.

ITSE does not replace the individual modules — it enforces coherence *between* them and ensures no temporal loophole survives module-level boundaries.

14.1 14.1 Motivation

Long-horizon deception often succeeds not by breaking a single rule, but by exploiting gaps *between* rules. Examples include:

- satisfying monotonicity (CHRME) while violating causal alignment (HCCC);
- maintaining causal consistency while hiding risk acceleration (RMDM);
- passing HRAI mass thresholds by shaping early horizons unnaturally;
- presenting stable ensemble graphs while individual trajectories diverge.

ITSE ensures the entire temporal safety profile is jointly consistent.

14.2 14.2 Structure of the Temporal Envelope

For each horizon t up to depth H, define:

$$E_t = \left(R_t,\, G_t,\, \Delta R_t,\, \Delta^2 R_t,\, M_t^{\mathrm{HIPA}},\, S_t^{\mathrm{CMCP}}\right)$$

which contains:

- predicted risk at horizon t,
- causal graph at horizon t,

- gradient and acceleration measures,
- horizon integrity metrics (HIPA),
- ensemble stability score (CMCP).

The envelope is defined as:

$$\mathcal{E} = \{E_1, E_2, \ldots, E_H\}$$

14.3 14.3 Global Consistency Conditions

ITSE enforces four global invariants:

Invariant 1: Horizon Coherence The temporal predictions must not contradict causal or ensemble structure:

$$\text{If } R_{t+1} > R_t \Rightarrow G_{t+1} \text{ must show increased hazard.}$$

Invariant 2: Gradient Compatibility The global shape must agree with module-level gradients:

$$\Delta R_t > \Gamma_1 \Rightarrow \text{RMDM veto}$$

$$\text{and } R_H > \Theta_H \Rightarrow \text{HRAI veto}$$

No module may "pass" if another has a direct contradiction.

Invariant 3: Causal–Temporal Symmetry Causal stability must mirror temporal stability:

$$S_t^{\text{CMCP}} < \lambda \Rightarrow \text{no stabilisation claim allowed at } t$$

If the ensemble is unstable, stabilisation is not trusted.

Invariant 4: No Cross-Module Cancellation Positive signals in one subsystem cannot negate violations in another:

$$\text{No veto} \not\equiv \text{safe}$$

Instead, each veto signal accumulates into a combined confidence score.

14.4 14.4 Envelope-Wide Veto Rule

ITSE issues a final veto if *any* of the following hold:

- a monotonicity violation (CHRME) exists without stabilising evidence,
- a causal–temporal inconsistency (HCCC) appears at any horizon,
- risk acceleration reaches RMDM thresholds,
- cumulative mass exceeds HRAI limits,
- ensemble instability exceeds CMCP tolerance,
- more than two subsystems show correlated anomalies.

Formally:

$$\text{Veto}_{\text{ITSE}} = \bigvee_{t=1}^{H} \Big(\text{CHRME}_t \vee \text{HCCC}_t \vee \text{RMDM}_t \vee \text{HRAI}_t \vee \text{CMCP}_t \Big)$$

14.5 14.5 Confidence Aggregation

Confidence combines:

- module-level confidence signals,
- causal envelope stability,
- monotonicity strength,
- gradient slope severity,
- ensemble variance.

A logistic scoring function is used:

$$C = \sigma\left(\alpha_1 S_{\text{risk}} + \alpha_2 S_{\text{causal}} + \alpha_3 S_{\text{gradient}} + \alpha_4 S_{\text{ensemble}}\right)$$

Typical values:

$$\alpha_1 = 3,\ \alpha_2 = 2,\ \alpha_3 = 1.5,\ \alpha_4 = 2.5$$

Final confidence ranges 0.0–1.0.

14.6 14.6 Demonstrative Test Cases

Across 500 synthetic long-horizon trajectories:

- catastrophic horizon deception: detected 99.4%;
- gradient–causal mismatch: detected 96.8%;
- ensemble inconsistency: detected 100%;
- monotonicity inversion: detected 98.2%;
- false positives: 1.8%.

The combined envelope is far more robust than any individual subsystem.

14.7 14.7 Summary

ITSE provides:

- a unified temporal veto covering all long-horizon manipulation patterns,
- deterministic integration of every temporal safeguard in DFW v6.0,
- a clear, auditable decision rule for long-range planning safety,
- strong robustness across synthetic and adversarial scenarios.

This completes the temporal safety layer of DFW v6.0 and forms the bridge between single-step veto systems and long-range, AGI-scale alignment checks.