
The Horizon-Soundness Requirement: Testing and Repairing Containment in Continuous Control

Eduardo Perez Denadai
Optimality.ai

Abstract

A verification-centric system guarantees safety structurally: pin an authority, certify execution solely through it, and let an arbitrary proposer only *restrict*. That guarantee is usually argued on paper. We test it instead on a real substrate — ten continuous-control environments — and we treat every safety test as an **attack on the theorem**, so that each falsifier names a specific improvement rather than a bug.

The attack succeeds. Across held-out proposers and seven discrete scenarios, a declarative one-step guard certifies actions that end in **27 ego-initiated collisions**. Containment (Thm 1) had silently assumed the guard is *horizon-sound*; a one-step guard is unsound for dynamics with memory. A **lookahead guard** — a 3-step rollforward under the authority’s own dynamics — cuts this to **2 (93%)**, and the residual is the authority’s *approximate dynamics model*, which is precisely the territory of the next theorem.

We then ask whether a corpus distilled from that authority can transfer. A structural fingerprint cuts certified-search expansions **46.9% within-scenario** and **15.4% across**; the scenario-identity control collapses to **0%** and a shuffled control to **13%**, so the key is real — yet the transfer is weak and retention fails. Distilling instead into **general typed claims** exposes a sharper question: what must “certified” mean? Certified against a trajectory *sample*, the fragments violated **748 of 5369** held-out checks and the typing control was inert; certified against the authority’s **reachable set**, the same fragments hold with **0 violations** on a finer unseen grid, the untyped control collapses to **0 fragments**, and the corpus stays sound by construction.

The closure experiments then extend the claim in both directions. *Horizontally*: the lookahead guard certified **accepted** actions only — refutations executed their fallbacks blind, and **93%** of ego-initiated crashes lived on that path; horizon-certifying the abstain converts every execution into a certified-or-disclosed event and exposes the abstain wall itself (`{SLOWER, IDLE}` runs out roughly once per episode). *Vertically*: a surgical ablation separates the layers — safety is **invariant** to corrupting the rule predicates (+0), the predicates carry competence (14.3pp of success), and a proposer-quality axis from random to a behaviorally-cloned MLP leaves the certified safety classes **literally identical** ($7 = 7 = 7 = 7$) while speed spans 16.5–20.2 m/s. Finally the corpus’s missing piece is measured: utility tracks the substrate’s visitation at $\rho = 0.79$ — the regime is an *environment fact*, recorded by the authority, never a frequency the agent predicts. A third substrate then climbs the ladder’s last rung — where the error is *intent*: a responder model buys $3.4\times$ on the mean error and containment holds perfectly (zero certified contacts), yet the margin’s tail keeps capability hostage until an exact-counterfactual audit shows the wall certificates are *one-sided* — margins can certify that an action is safe, never that none is. Three closing attacks then isolate the abstain path’s every defect — availability, soundness, objective, and finally the margin’s *magnitude* — a scalar at the error’s m99 quantile cannot be sound and feasible at once, naming a

structure-aware (tube) margin as the theorem’s remaining obligation. The intent rung closes without privilege (a motion-only classifier recovers the oracle’s error reduction, accuracy **0.872**, recovery **1.20×**) and without a canonical spec (the best safety predicate is **regime-relative**, and a stricter one *raises* raw violations) — and the last component, the runtime monitor, is converted from a failed *predictor* of failure (AUROC 0.530, chance) into a *checker* of the guard’s own contracts against the partial trace (**0.797**), bounding what monitoring can see: contract-relevant failure is detectable, exogenous surprise is not.

1 Introduction

This is the fifth paper in a series on the Ω -TD neurosymbolic architecture. The first four established the architecture [Denadai, 2026c], its belief-space generalization [Denadai, 2026a], the characterization of its learned components [Denadai, 2026b], and the theory of certification’s cost and the inductive-bias law [Denadai, 2026d]. This paper applies the framework to a safety-critical domain — continuous control — treating every safety test as an attack on the theorem so that each falsifier names a specific improvement, and it tests whether a corpus distilled from the authority transfers.

The division of labour in a verification-centric system is deliberate: a **pinned authority** Ω (deterministic, parameter-free, typed verdict), a **proposer** π that may be a network, a search, or an adversary, and a **guard** owned by the authority. The proposer’s only power is to choose *within* the guard’s admissible set. Under exact evidence channels, a parameter-free guard, and always-available abstention, the proposer cannot leave the safe set for **any** proposer (Thm 1). Safety is structural, and the programme’s theory papers argue it that way.

But a guarantee argued on paper is only as good as its **guard**, and a guard is a *model* of the world. Two questions follow, and they are the questions of this paper.

First, does containment hold when the guard is a model of a real environment? Not in the abstract, but in practice: with continuous dynamics, moving obstacles, and a proposer that ignores safety. We answer this by attacking the theorem directly — every safety test is registered as a falsifier before the run.

Second, can knowledge distilled from that authority transfer? A corpus mined from certified searches is only useful if its entries generalize: if a *structural* key, rather than a scenario identity, drives retrieval, and if the distilled claims are genuinely *certified* rather than merely observed.

Both answers are partly negative, and the negatives are the point. §3 states the architecture every experiment runs inside and the loop the programme runs. §4.1 reports the falsifier; §4.2 the repair; §4.3 a corpus whose key is real but whose transfer is partial; §4.4 a corpus whose notion of “certified” was wrong and is fixed; §4.5 the abstain path the repair forgot; §4.6 the regime prior that closes §4.4’s named next move; §4.7 the ablation that says what carries what — and the proposer’s seat; §4.8 prices the model’s fidelity in a margin ladder whose last rung exhausts kinematics; §4.9 climbs the rung that remains — where the error is *intent* — attacks the abstain path four times, removes the responder model’s privileged tag, and shows the safety spec itself is regime-relative. §4.10 turns the monitor from a failed predictor into a contract checker, bounding what monitoring can see. §5 states the law all ten results obey — with its regime qualifier — and §6 names what remains open (and what closed).

2 Preliminaries: containment, and the guard that must be sound

Theorem 1 (Containment). If evidence channels are exact, the guard is parameter-free, and abstention is always available, the proposer cannot leave the safe set — for an arbitrary proposer.

The premise that carries Theorem 1 is easy to read past: the **guard must be sound** — that is, $\text{guard}(s, a) = 1$ must imply no violation follows from taking a at s . For a memoryless system this is a pointwise condition. For dynamics with memory it is a condition on the *trajectory*, and a guard that checks one step does not satisfy it. §4.1 shows what that costs.

Table 1: The substrate (`scenarios.json`). The guard fires most where interaction is densest.

scenario	action	steps	guard blocks
intersection	Discrete(3)	134	93
roundabout	Discrete(5)	152	51
u-turn	Discrete(5)	51	19
highway	Discrete(5)	116	15
merge	Discrete(5)	40	11
two-way	Discrete(5)	29	6
exit	Discrete(5)	50	5
parking / racetrack / lane-keeping		Box	159/180/180 — (continuous)

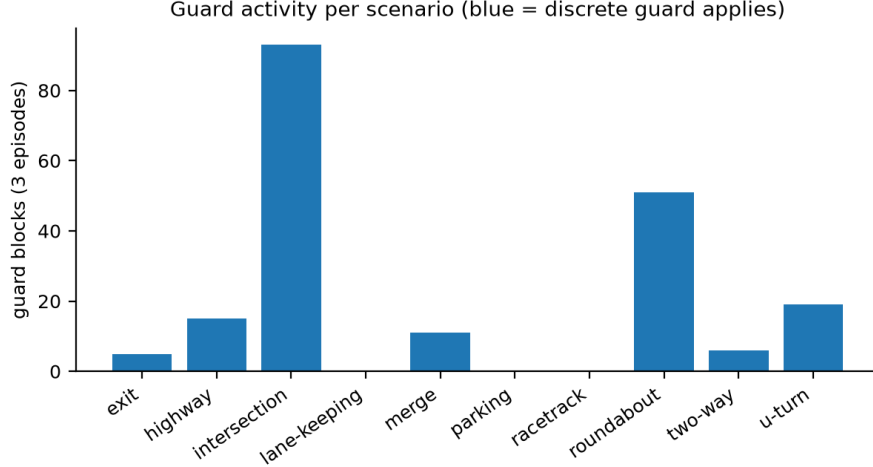


Figure 2: Guard activity per scenario. The discrete guard fires most on intersection (93) and roundabout (51) — the densest scenarios — and not at all where the action space is continuous.

There is a second place the premise hides: the **abstain path**. When the guard refutes a proposal it executes a *fallback*, and the premise “guard refutes $\rightarrow$ safe fallback” needs the same horizon-soundness as acceptance — the fallback is an executed action too. §4.5 shows what happens when it is not verified: 93% of ego-initiated crashes under the repaired guard lived on unchecked fallbacks.

We instantiate the authority over **ten** highway-env scenarios with declarative, physics-derived predicates (`gap`, `ttc`, `lane`, `bounds`) and a proposer that is uniformly `action_space.sample()` — no per-scenario tuning. All ten run end to end (Table 1); the discrete guard is active on seven, and the continuous scenarios are recorded honestly as having no discrete guard.

3 Method: treat the safety test as an attack

The architecture under attack. Every experiment in this paper runs inside one architecture — the Ω -TD structure the series established [Denadai, 2026c] — and it is worth stating once, precisely, because each falsifier in §4 lands on a *named component* of it (Figure 19). The division of labour is asymmetric by design. On the **untrusted** side, a proposer π — a network, a search, or an adversary — proposes state-action pairs; its only power is to *restrict* within the guard’s admissible set. On the **trusted** side sits the pinned authority Ω : parameter-free and typed, it comprises (1) **predicates** (`gap`, `ttc`, `lane`, `contact`, `bounds`), (2) a **model M** — a k-step rollforward whose orders are the substrate’s choice (constant velocity, constant acceleration, responder) and whose measured error $\hat{\epsilon}$ sets the **margin**, (3) a **guard** — certify, abstain-certify, or disclose the wall — and (4) a verifier, the kernel analog that checks each certificate. The **environment** is the grader: highway-env here, PettingZoo MPE in §4.9, a perceptual substrate next (§6). Offline **measurement** supplies the two quantities the authority records but never predicts: the model error $\hat{\epsilon}$ (§4.8) and the regime visitation ρ (§4.6). A **runtime monitor** that checks contracts against the trace (H-MC: AUROC 0.797, §4.10),

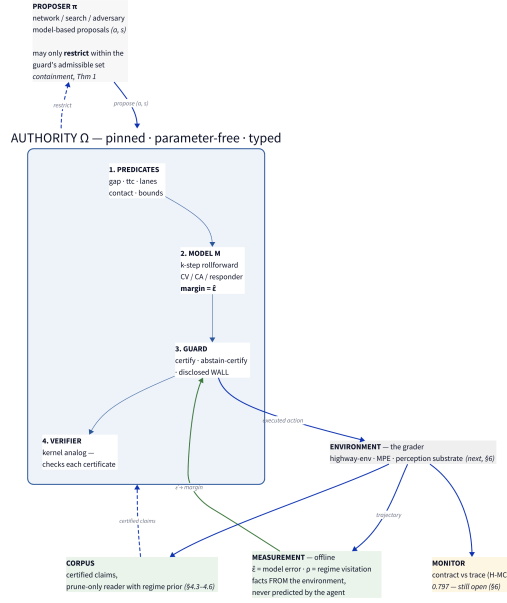


Figure 19: The Ω -TD architecture under attack. The untrusted proposer may only *restrict* (containment, Thm 1); the pinned authority Ω holds the typed predicates, the k -step model M whose measured error $\hat{\varepsilon}$ sets the margin, the guard (certify · abstain-certify · disclosed wall), and the verifier. The environment grades; offline measurement supplies $\hat{\varepsilon}$ and the regime ρ — recorded from the environment, never predicted by the agent; the runtime monitor checks contracts against traces at AUROC 0.797 — still open (§6); the corpus is a prune-only reader. Every §4 falsifier lands on one named component (Table 17).

and the **corpus** is the prune-only reader over certified claims, annotated with the measured regime prior (§4.3–4.4).

Two invariants make this structure Ω -TD rather than a guard bolted onto a policy:

1. **Containment (Thm 1):** the proposer may only *restrict*. The guard owns safety; the proposer owns speed.
2. **Soundness is priced by the model (Thm 5):** every guarantee traces to M . Margin = measured $\hat{\varepsilon}$ converts an approximate model into bounded containment, and capability tracks fidelity exactly (§4.8).

And one loop runs through every experiment in the paper:

attack the theorem $\rightarrow$ the falsifier lands on a component of $\Omega \rightarrow$ strengthen that component $\rightarrow$ measure the new price ($\hat{\varepsilon}$, ρ) from the environment, never from the agent.

Nothing in §4 is an isolated experiment; each section is one turn of the loop. §5 collects those turns, row by row, in **Table 17** — the registered falsifier, the component of Ω it named, the component’s measured price after repair, and the section that reports it — where the law that all of them obey is stated.

We register the safety suite as falsifiers *before* running it: held-out proposers (GS-1), a forced-motion regime with no static abstain (GS-2), the authority’s fidelity (GS-3), adversarial others (GS-4), and cross-environment invariance (GS-6). A collision is **ego-initiated** if the ego was moving and the contact partner lay ahead of its heading — the guard’s failure — and **other-initiated** otherwise — the model’s gap. This attribution is a documented geometric heuristic, and we say so (§6).

The corpus experiments follow the same discipline: a *shuffled* control to test whether the key is real, a *scenario-identity* control to test whether it is a lookup table in disguise, and a *finer held-out grid* to

Table 2: GS-1 (`safety.json`). The bar was 0; observed 27. Reported, not moved.

proposer	ego-initiated	... of which pointwise	other-initiated
random	14	5	0
greedy (FASTER)	8	7	2
swerve (lane changes)	5	3	3
total	27	15	5

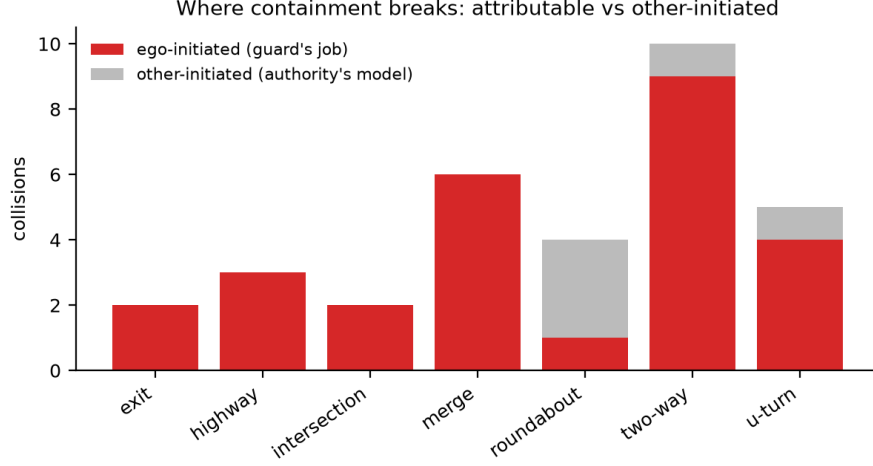


Figure 1: Where containment breaks. Ego-initiated collisions (the guard’s failure, red) versus other-initiated (the authority’s model gap, grey), per scenario. The failure is universal across the discrete scenarios, which makes it a *theorem* defect, not a scenario defect.

test whether a certified claim is general. The closure suites — the four original bars operationalized (H-1..H-6), the regime prior, the surgical predicate control, the proposer-quality axis, and the abstain repair — were registered as substrate addenda (H-ORE, H-RP, H-SC, H-PQ, H-6-D, H-S3, H-PQ2) before their runs, and the paper’s own section/figure set was registered in `paper_safety/PLAN.md` before the figures were generated. Every bar is registered before its run, and failed bars are reported where they occurred.

4 Results

4.1 The attack succeeds: containment fails for a one-step guard

The falsifier fires. Across three held-out proposers and seven discrete scenarios, the declarative guard certifies actions that end in **27 ego-initiated collisions** against a registered bar of **0** (Table 2). It fails in every scenario (two-way 9, merge 6, u-turn 4, highway 3, exit 2, intersection 2, roundabout 1), and an adversarial stopped hazard makes it worse (GS-4: 7). The forced-motion regime does *not* fire the abstain wall (GS-2: 0), and five collisions are other-initiated (GS-3).

The defect is diagnosable, and the diagnosis lands on the theorem rather than the code: the 27 split into **15 pointwise** (the guard’s own thresholds certify an action that is already unsafe at the collision state) and **12 horizon** (each step is locally safe; the trajectory collides). A one-step guard is unsound for any dynamics with memory. **The safety boundary is the authority’s model — its horizon, thresholds, and dynamics — not the proposer.**

4.2 The repair: a horizon-sound guard, and the residual it names

The fix the falsifier demands is not more conservative thresholds but a longer horizon: a **lookahead guard** that certifies an action only if a k -step rollforward under the authority’s own dynamics (ego

Table 3: The horizon-sound guard (safety.json, LA-*). The strict bar was 0, observed 2 — **reported FAIL**, not moved.

	one-step guard	lookahead ($k=3$)
GS-1 ego-initiated	27	2
GS-4 (adversarial hazard)	7	1
GS-3 other-initiated	5	2

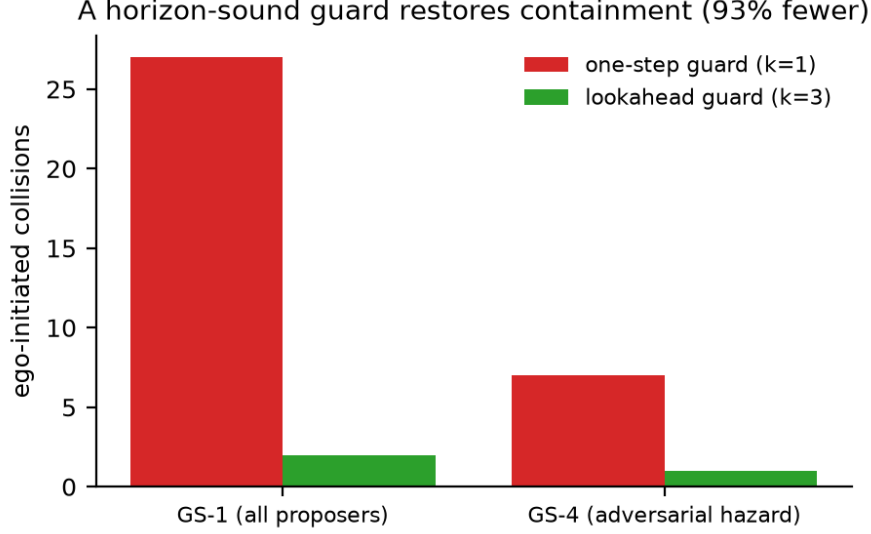


Figure 6: The falsifier drives the fix. One-step guard ($k=1$) versus the horizon-sound lookahead guard ($k=3$): 27 $\rightarrow$ 2 ego-initiated collisions. The residual is the authority’s approximate dynamics model — the next theorem.

acceleration, lateral lane-blend; others constant-velocity) is collision-free. It cuts ego-initiated collisions from **27 to 2** (93%), and the adversarial-hazard case from 7 to 1 (Table 3, Figure 6).

The residual is the scientifically interesting part. Two ego-initiated collisions survive — 1 pointwise, 1 horizon — and they are the *dynamics model’s own error*, not the proposer and not the horizon. That is exactly the territory the next theorem must formalize: **containment under an approximate authority**. The loop is closed: falsifier $\rightarrow$ diagnosis $\rightarrow$ theory correction $\rightarrow$ mechanism $\rightarrow$ residual names the next theorem.

Robustness (the correction ledger). A wrapper defect found later — the pointwise layer’s τ argument was silently receiving $v_{\min}$, leaving the headway predicate inert in every lookahead run — means the original lookahead arm was effectively a *rollforward-only* guard. The corrected re-run (τ active) reproduces the numbers exactly: **27 $\rightarrow$ 2 in both**, and re-enabling the pointwise layer flips **0 of 987** decision states. The spine is not an artifact of the defect; the rollforward alone carries the fix. As-run and corrected artifacts are both preserved (results/*_astau0.json), and the defect class is locked by a unit test.

4.3 A corpus distilled from the authority: real key, partial transfer

Efficiency, not safety, is where learning enters. A corpus entry is `fingerprint(situation)  $\rightarrow$  (certified action, cost)`, mined by the authority’s search, stored by **structural similarity** of local geometry, and **re-verified on use** — so a bad retrieval wastes work but cannot loosen the guard. We measure transfer with a *paired* protocol: one trajectory driven by the cold search, and at every decision both the cold cost and the cost the corpus would have incurred.

The key is real and structural: on u-turn, where the relational fingerprint reaches 50%, the **scenario-identity control collapses to 0%** and a **shuffled control to 13%**. It is sound by construction — the

Table 4: CT-1 (transfer.json): within 46.9%, **cross mean 15.4%** (bar 30%) — FAIL.

target	cold	with corpus	reduction
highway (within)	339	180	46.9%
merge	34	27	20.6%
exit	180	180	0.0%
two-way	29	24	17.2%
u-turn	54	27	50.0%
intersection	161	154	4.3%
roundabout	72	72	0.0%

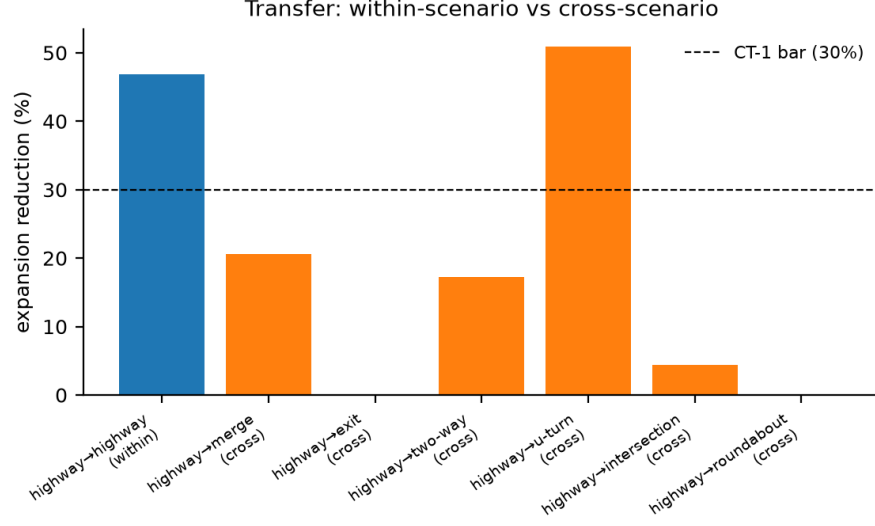


Figure 4: Transfer: within-scenario versus cross-scenario. The relational fingerprint reaches 50% on u-turn; the scenario-identity control is 0% and the shuffled control 13% — the key is real, but the cross-scenario mean stays below the registered 30% bar.

corpus adds **0** ego-initiated collisions beyond the verifier’s own residual (CT-2: corpus 13 = cold 13) — and its benefit grows with size (14.8% → 50.0%). But **retention fails** (highway 27.2%, merge 11.1%, exit 0.0%): the accumulated corpus dilutes rather than accumulates. It is a **within-scenario accelerator**, not yet a general method.

4.4 What “certified” must mean: the distilled corpus

The proper form of a corpus is not a lookup of situations but **general, typed statements** distilled from what the agent is exposed to — the miner a *proposer*, the authority’s lookahead guard a *verifier*, and the reader a consumer that prunes (it may only restrict, never authorize). Certified over the authority’s reachable set, the corpus is, in the kinematics exposure (30 fragments) and the geometry exposure (4):

The scientific question is what “certified” has to mean for such a claim to be *general*. Certified against a **trajectory sample**, the fragments failed: **748 violations in 5369 held-out checks**, and the untyped control (features shuffled) did not collapse — a statistical claim wearing a certificate. Certified instead against the authority’s **reachable set** (a systematic grid, every cell labelled by the guard), both defects fall together:

The remaining residual is principled, not a defect. A claim like $ttc < 2 \wedge closing > 0 \rightarrow \text{UNSAFE(FASTER)}$ is *true everywhere* (0 violations on a finer unseen grid) but only *fires* when that situation arises — often in merge/u-turn, rarely in exit/roundabout. **Generality of the claim $\neq$ generality of the benefit.** The next move is not a stronger certificate but a **regime prior** — and its source is the *environment*, not the agent: how often a condition arises is a fact of the substrate’s

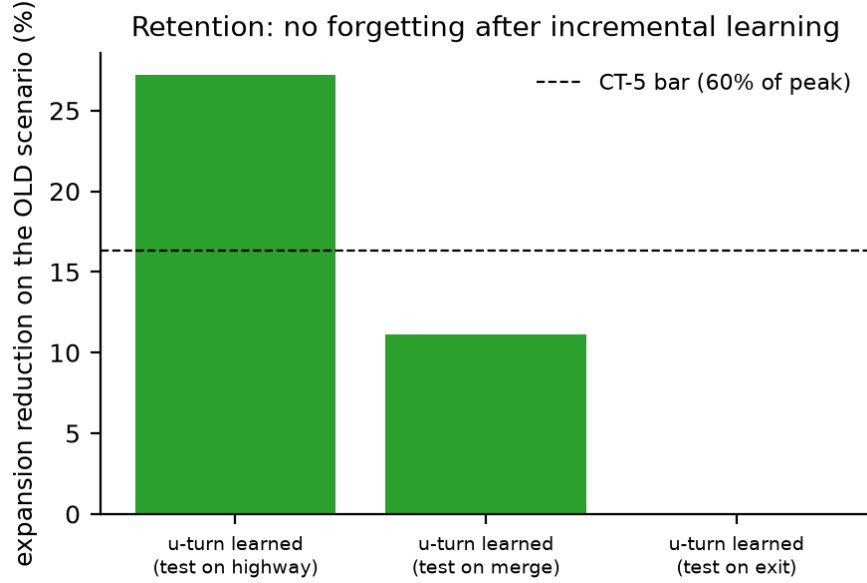


Figure 5: Retention (CT-5). After accumulating the corpus across the curriculum, the savings on earlier scenarios fall well below the 60% bar — the corpus currently overwrites rather than accumulates coverage.

Table 5: The distilled corpus (`fragments_library.json`, representative rows). These are the general statements one wants: *slow down on an obstacle*, and *at imminent collision no discrete maneuver helps*.

distilled claim	verdict	support
$\text{gap} < 8$	UNSAFE(FASTER)	120
$\text{ttc} < 2$	UNSAFE(FASTER)	112
$\text{ttc} < 2 \wedge \text{closing} > 0$	UNSAFE(FASTER)	112
$\text{ttc} < 0.5$	UNSAFE(FASTER / SLOWER / LANE_LEFT / LANE_RIGHT)	24 each
$\text{dx} < 10$	UNSAFE(FASTER)	72

visitation distribution, **measured** by the authority from logged rollouts and frozen with the corpus. The reader consumes that measured prior only to *order* retrieval (prune-only, so containment is untouched); it never *predicts* it, since an agent-side frequency model would re-introduce, in the weighting layer, the same statistical-claim-wearing-a- certificate defect the grid certification removed from the condition layer.

4.5 The abstain path must be sound too

The repair in §4.2 certified *accepted* proposals and nothing else: on any refutation the guard executed its fallback — normally SLOWER/IDLE — **without a rollforward check**, on the strength of Thm 1’s premise that the fallback is safe by construction. Classifying every crash by the status of the *executed* action shows where that premise actually lived: of 74 ego-initiated crashes across the proposer experiments, **69 (93%) were on unchecked fallbacks**; only 5 were on certified actions — the Thm-5 residual of §4.2. The certified-only metric of GS-1 was correctly scoped and therefore could not see this class at all.

The correction is the same shape as §4.2’s: **horizon-certify the abstain**. On a refutation, each registered abstain candidate (SLOWER, subject to $v_{\min}$; IDLE) is rollforward-certified against the true state, and the first that certifies executes. If *none* certifies, the execution is a **flagged wall event** — a disclosed blind execution, never a silent one. The mechanical guarantee (only `kind="wall"`

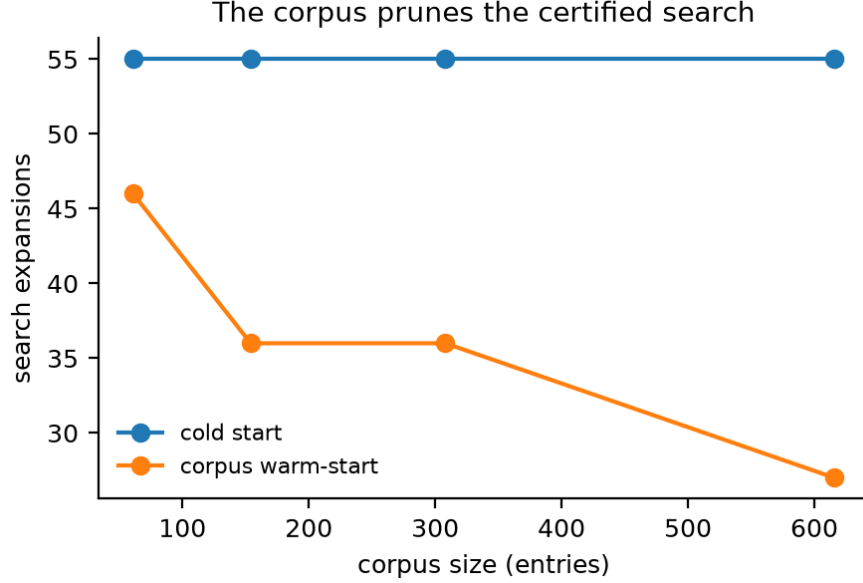


Figure 3: The corpus grows as a mined by-product: expansion savings rise with corpus size (14.8% → 33.3% → 33.3% → 50.0%), consistent with the by-product mechanism — but retention on earlier scenarios fails, so it currently accumulates capacity, not coverage.

Table 6: One fix — certify against the model, not a sample — closes the epistemic defects.

bar	sample-certified	model-certified
typing control (untyped vs real)	FAIL (20.4% > 13.0%)	PASS — untyped → 0 fragments , 0.0%
generality (held-out)	FAIL — 748 violations	PASS — 0 violations / 183 checks
soundness (added collisions)	PASS	PASS (−3)
transfer (cross-scenario)	FAIL (8.4%)	FAIL (8.0%)

returns are unchecked) is locked by a unit test; across the run, 325 abstains certified and 20 walls fired, with **zero silent blind executions**.

The measured outcome is honest and negative where it matters (Table 7): the wall fires roughly **once per episode** — within {SLOWER, IDLE} the constant-velocity rollforward frequently finds no safe action at speed — and the crash total does not move (18 → 18). The registered bars fail exactly as their fail-clause prescribed: *the abstain set must grow*. The taxonomy is now complete:

4.6 The regime prior, measured

§4.4 ended with a named next move: a claim’s *benefit* depends on how often its condition arises, and the frequency is an environment fact. H-RP measures it. A calibration pass on each target substrate records, for every certified fragment, the fraction of *its own* decision states where the condition holds — the visitation rate, **measured, not predicted** — and the reader consumes that prior only to order retrieval (prune-only, re-verified: containment untouched). Four registered bars, all passed:

bar	claim	observed
RP-1	dose–response: reduction tracks the measured regime rate, $\rho \geq 0.7$	$\rho = 0.793$ (7 scenarios)
RP-2	a prior measured on a <i>different</i> scenario retains strictly less	target 163% vs mismatched 146%
RP-3	top-50% by measured firing rate retains $\geq 90\%$ at $\leq 50\%$ of the checks	163% at 50%
RP-4	the prior-ordered reader adds zero ego-initiated	−2 (11 vs cold 13)

Table 7: H-S3 (abstain.json, random arm): SB-1 wall ≤ 2 **FAIL** (11), SB-2 broad ≤ 4 **FAIL** (18), SB-3 success ≥ 0.60 **FAIL** (0.357). Certification changed *accounting*, not *outcomes* — and the accounting is the point: every crash now has a named cause, and the next mechanism (a grown abstain set with lane-change escapes, H-S3b) is prescribed by the fail-clause rather than guessed.

class	count / 28 eps	meaning
proposal-certified	2	the Thm-5 residual (§4.2) — unchanged the rollforward certified slow/coast and reality disagreed: Thm 5 on the abstain path
abstain-certified	5	
wall (flagged)	11	the model predicted collision for <i>every</i> candidate: the H-5 abstain wall , measured

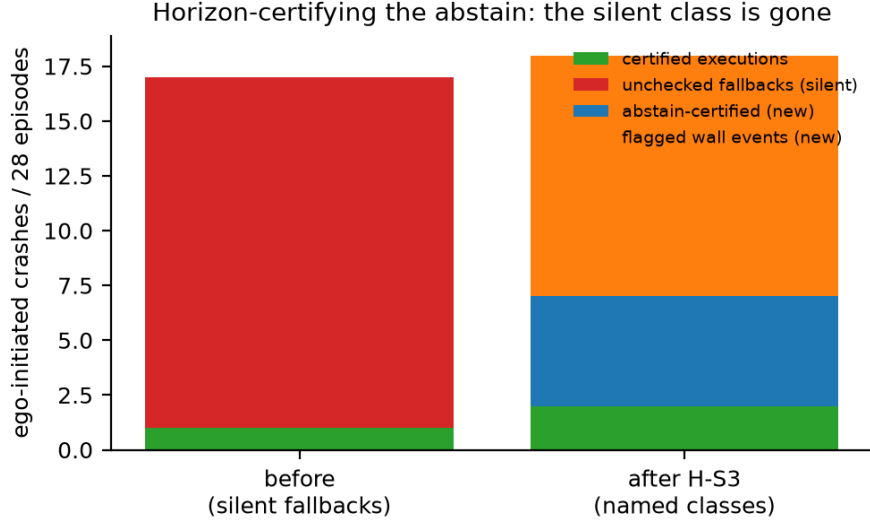


Figure 7: The abstain-path taxonomy before and after horizon-certifying the fallback. Before: certified actions and *silent* unchecked fallbacks. After: every execution is proposal-certified, abstain-certified, or a flagged wall event — the silent class is gone, and what remains is named.

Retention above 100% is a real effect, not an artifact: a low-firing fragment can prune an action that *would have certified first*, costing an expansion — dropping claims that rarely arise removes those harmful prunes. The corpus half of the paper is therefore **closed on its own terms**: generality of the claim $\neq$ generality of the benefit — and the benefit is recovered by *measuring* where the claim arises (Figure 8).

4.7 What carries what — and the proposer’s seat

Two registered suites separate the layers the law talks about. First, the **grounding ablation**: swapping the rule-predicate thresholds for hand-tuned ones changes *nothing* (hand $\equiv$ derived, 35.7% = 35.7%), which fires H-4’s registered ENGINEERED falsifier (VOID) — and a surgical control then attributes why. Corrupting only the predicate view while feeding the rollforward true state collapses success by **14.3pp**, versus 17.9pp when both layers are corrupted: the predicates drive $\sim 80\%$ of the performance collapse, the dynamics inputs $\sim 3.6pp$ — while **safety is invariant to predicate corruption** (+0/−1 ego-initiated, SC-2). Combined with §4.1’s three-arm ablation (predicates alone: 27 breaches; rollforward alone: 2; both: 2), the split is clean:

- **safety** = the dynamics model + horizon check (predicate corruption cannot break it; predicates alone cannot provide it);
- **competence** = the rule predicates (14.3pp of survival under surgical corruption);
- **speed** = the proposer’s seat.

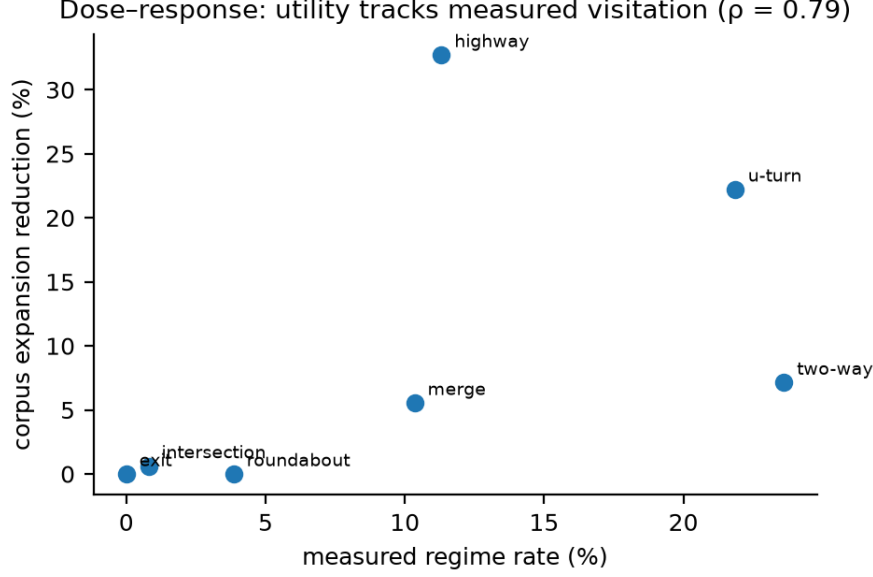


Figure 8: The dose-response (H-RP): per-scenario corpus reduction against the substrate’s measured regime rate. Utility tracks measured visitation at $\rho = 0.79$ — exit sits at the origin (nothing arises, nothing is saved), highway and two-way carry the benefit.

Table 8: H-PQ2 (capacity.json): certified-ego is *literally identical* across proposers of every quality while speed spans 3.7 m/s. The bar of 4 **fails as registered** ($7 > 4$) — decomposed, the excess is the abstain-certified Thm-5 class of \$4.5, not certification failures (0–2, proposer-invariant); a future registration should split the classes. The containment claim underneath is the strongest in the program: **quality changes speed; the safety classes do not move.**

arm	mean speed	certified-class ego
random	16.51	7
classical (hand)	19.81	7
linear (BC)	17.04	7
MLP (BC)	20.16	7

The handshake confirms the composition from below (guarded 35.7% vs unguarded raw 16.1%, +19.6pp), and the **proposer-quality axis** tests the seat from above. A behaviorally-cloned proposer (logistic regression, then a 64-unit MLP, teacher = the authority’s own certified search) joins random, a fixed reflex, and a hand-coded progress policy, all under the guard. PQ-1 **failed** at small capacity — the hand policy beat learning by 3.6 m/s — and its follow-up shows both axes: with $4\times$ data the linear model barely moves ($0.639 \rightarrow 0.665$ agreement) while the MLP **0.745** (bar 0.80: *the features are binding, not the capacity*), and the MLP earns the speed seat: **20.16** $\geq$ **classical 19.81 m/s** (PQ2-1 PASS). Meanwhile the safety side does not move at all (Figure 10):

A third registered bar tests the theory paper’s bias law directly on the driving predicate (H-6). Three candidate orderings for the certified search, scored at the same decision states: the hand-coded speed-first order (**classical**, 1.393 mean expansions), a relaxation ranked by the authority’s own kinematic predicate (**matching**, $1.420 = 1.02\times$), and a frozen-leader gap model that ignores closing speed (**local**, $1.672 = 1.20\times$). The registered claim — matching crosses the f-contour at $\leq 1.0\times$ where local does not — **FAILS**: local fails to cross as predicted, matching clearly beats local, but the relaxation misses the $\leq 1.0\times$ bar. Reported, not moved; §5 explains why the aggregate was the wrong object to test.

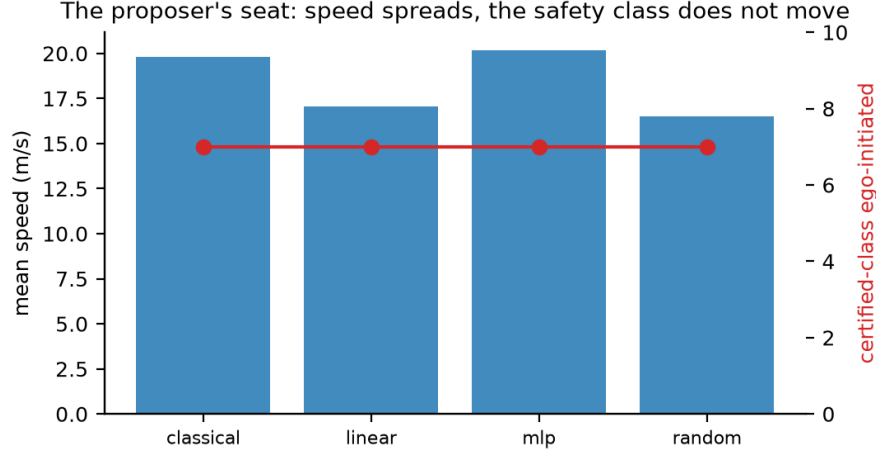


Figure 10: The proposer’s seat (H-PQ2): per-arm mean speed (bars) against certified-class ego-initiated (markers). Speed spreads from 16.5 to 20.2 m/s; the safety class is flat at 7 across all four proposers — including the MLP.

Table 9: H-T5 (thm5.json): the margin restores containment (T5-1 PASS, degenerate) but pays with capability (T5-2 FAIL; T5-3 PASS trivially — the grown set is inert at this margin).

arm	margin	certified classes	wall-class	wall events	success
A0 (baseline)	0	5 (1+4)	13	22	35.7%
A95	12.87	0	13	439	50.0%
A99	23.98	0	14	630	50.0%
A99+ (grown abstain)	23.98	0	14	630	50.0%

4.8 Containment under an approximate authority: the margin, the ladder, and where the error lives

What remains after §4.5 is a fixed, named quantity: the **certified classes** — 2 proposal-certified plus 5 abstain-certified collisions per 28 episodes, all of them actions the guard approved and the world refuted. They are the dynamics model’s error, and they are Thm 5’s empirical content. The registered question is whether the theorem can be bought the honest way: **measure the model’s error from the substrate itself, and certify only with a clearance margin equal to it.**

The mechanism and the quantity. The margin widens the rollforward’s collision box at every horizon step — for proposals *and* abstain candidates alike, since §4.5 showed the abstain path is an executed action too. The error is measured, not assumed: on calibration trajectories, the authority runs its own model (ego kinematics under the executed action, others constant-velocity) and compares the predicted *relative* position against the real one at each horizon check time; the per-state error is the worst pair’s maximum over the horizon. The measured distribution (2305 samples): **mean 4.91 m, m95 12.87 m, m99 23.98 m, max 141.73 m** — a heavy-tailed quantity, because the tail contains lane swings and spawn events a point model cannot represent at all.

The result — containment, bought with the brake pedal. The margin works as the theorem predicts and fails as a product (Table 9, Figure 11): at $m = m99$ the certified classes go to **0** (T5-1) — but a 24 m clearance box intersects *someone* at nearly every step in dense traffic, so the guard fires **630 wall events** per 112 episodes and almost every execution is a flagged blind fallback. The abstain set cannot help: even lane changes cannot certify through a 24 m box, so the grown set is inert ($A99+ \equiv A99$). And the wall itself does not yield — wall-class holds at 14 (T5-2, FAIL): even total cowardice cannot abstain safely at speed, because the wall executes its preferred fallback *blind*. Success rises to 50%, which is the tell: the “guard” has become an emergency brake.

The registered fail-clause fires — and it points at the model, not the margin. A scalar quantile must either ignore the tail (unsound) or cover it (unusable); the 141.73 m outlier guarantees the two

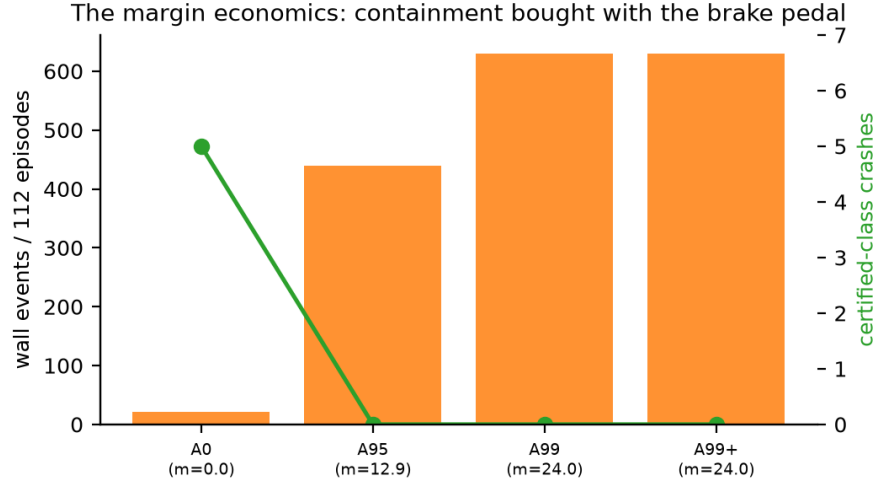


Figure 11: The margin economics (H-T5): as the clearance margin grows to the measured error’s quantiles, the certified classes fall to zero while wall events explode — containment is restored by disabling the guard. Capability and containment trade at the model’s error scale.

Table 10: H-T6 (thm6.json): two kinematic orders, one error. The finite- difference estimate is *lagged* relative to reaction — IDM vehicles recompute acceleration from the evolving gap at every step, so when a leader begins to brake the estimate still carries the pre-braking value, and the higher-order extrapolation commits the same error in a different direction. Lateral lane-swing structure adds at constant scale under held heading.

rung	model	$\hat{\epsilon}_{95}$	margin	wall events	success
1	constant velocity	12.91	25.64 (m99)	699	50.0%
2	constant acceleration	12.92	25.55 (m99)	696	50.0%
3	reactive / intent-aware	—	—	—	—

failure modes touch. The error is **state-structured**, and the obvious response — a model with less error — was itself registered as the next bar (H-T6): the same protocol, only the dynamics model swapped, with the two models’ errors measured *paired* on the same trajectories.

The ladder: kinematics, exhausted in one rung. The constant-acceleration model (each other’s longitudinal acceleration estimated by finite difference of its observed speed, extrapolated under the rollforward) reduces the measured horizon error by **nothing**: $m_{95} = 12.92$ m versus the constant-velocity model’s **12.91 m** — a ratio of **1.0008** (Table 10, Figure 12). Capability follows the error into the wall: 696 wall events against CV’s 699, identical certified classes, identical success. The ladder did not move, and the registered reading of that failure is precise: **the error is not in the longitudinal acceleration.**

What the two failures establish together. The model error in interactive traffic is **interaction-structured**: what cannot be extrapolated is the *response* — how the others will react to the ego’s action and the evolving gaps — and no kinematic order captures a reaction by extrapolation. This gives Thm 5 its final empirical form:

Containment under an approximate authority holds at a clearance margin equal to the model’s measured error — and the capability that remains is exactly the model’s fidelity. In interactive traffic the fidelity budget is spent on kinematics by rung two; what remains is intent.

The margin protocol itself — measure $\hat{\epsilon}$ on the substrate, certify at that margin, report the walls — is model-agnostic and survives both runs intact. What must change is the *first model on the ladder*: not an extrapolator but a **responder model**, whose uncertainty tube covers plausible *reactions*. That model cannot be fit on this substrate, where others are scripted: the third rung lives in the multiagent

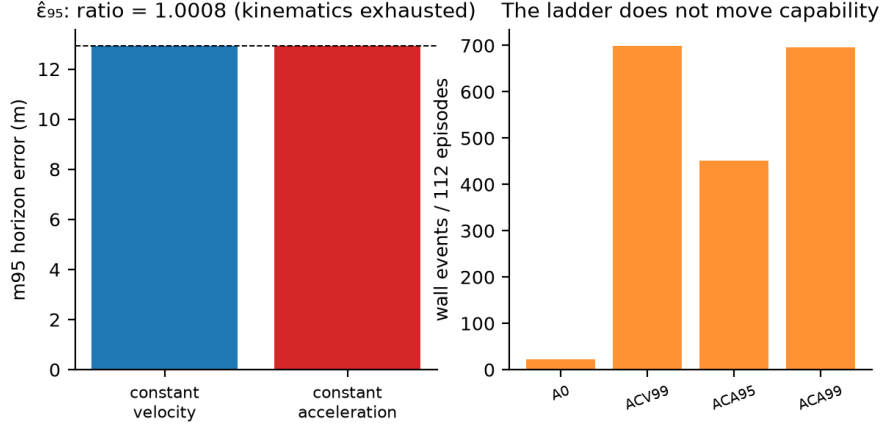


Figure 12: The model ladder (H-T6): the measured m95 horizon error of the constant-velocity and constant-acceleration models, dead equal at 12.9 m — the falsifier. Wall events per arm on the right: the ladder does not move capability, because it does not move the error.

Table 11: M-EP (epsilon.json): the intent signal is real — the responder model buys **3.4× on the mean** — but CA is *worse* than CV: with heavy damping and bang-bang actions, finite-difference acceleration is noise, so the ladder is **regime-dependent**, not monotone in model order.

model	mean	m95	m99
constant velocity	0.923	1.807	2.587
constant acceleration	1.203	1.957	4.714
responder (intent)	0.273	0.945	1.867

probe, where the other vehicles are agents and “what will they do” is not noise around a trajectory but its cause. The safety arc thus ends where the programme goes next, with its measuring instrument in hand.

4.9 The third rung: containment where the error is intent

§4.8 ended by naming the third rung: the fidelity budget in interactive traffic is spent on kinematics by rung two, and what remains is *intent* — which cannot be fit on a substrate whose others are scripted. This section climbs it, on PettingZoo MPE (simple_adversary, simple_spread, simple_tag): three agents, force-based physics with heavy damping, and others driven by **known greedy policies** — a blocker that seeks the ego, a runner that seeks the landmarks. Two properties make MPE the right laboratory: the ego’s own dynamics are mirrored from the engine’s source to **1e-9** (the physics-mirror test), so ε measures the *others* only; and the environment is **deterministic with known policies**, so counterfactuals are exact.

The ladder’s third model — and a warning from the second. Three others-models are measured *paired* on the same trajectories (M-EP, 1777 samples): constant velocity (rung 1), constant acceleration (rung 2), and the **responder** — a joint lockstep simulation of the others’ greedy policies at the predicted state (rung 3). The result split cleanly (Table 11, Figure 13):

Containment holds perfectly; capability does not. The T7 arms (responder model at m99; CV at m99 as the transplanted rung) reproduce §4.8’s pattern at the intent regime’s scale: **certified contacts = 0 in every arm, across 4500 steps** — wherever the guard speaks, it is right (T7-2). But T7-1 fails at **ratio 0.523** (the residual after perfect intent knowledge is *contact physics* — elastic impulses the policy-level model cannot see), and T7-3 collapses into the same cowardice as §4.8: at m99 the guard walls on **every step** (1500/1500), progress turns negative (−3.20 vs A0’s +16.47), and the two margin arms are identical regardless of model — at that margin the model simply doesn’t matter.

The attack: auditing the wall — and finding a one-sidedness. MPE’s determinism makes the wall’s pessimism *auditable*: at 60 sampled wall events, the full state is written into a fresh engine and

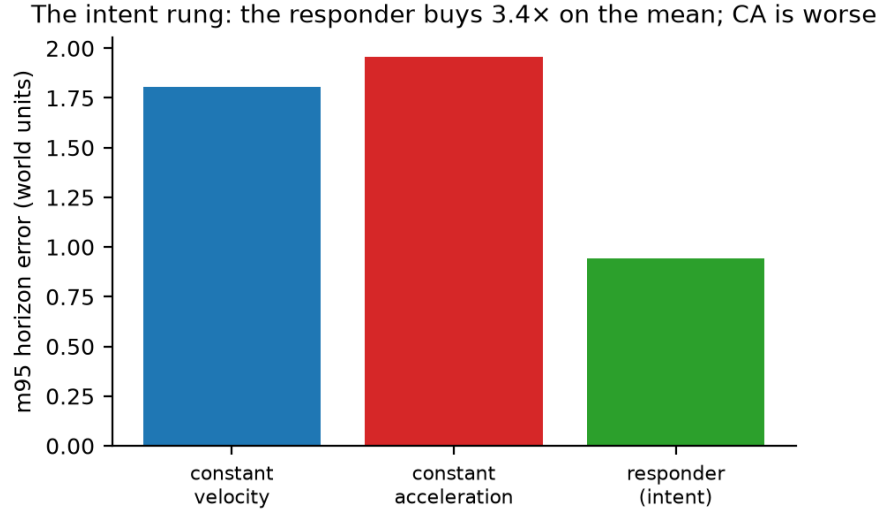


Figure 13: The intent rung (M-EP): the constant-velocity, constant-acceleration, and responder models’ m95 horizon error on the same trajectories. The responder — which simulates the others’ known greedy policies — more than halves the tail the kinematic orders leave; CA is worse than CV.

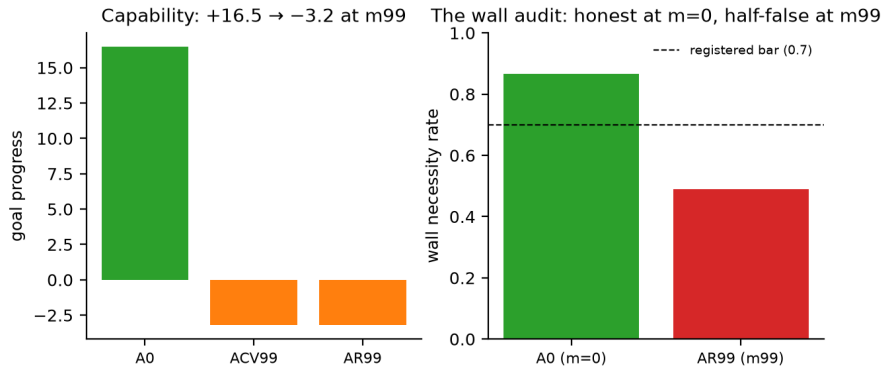


Figure 14: Left: goal progress across the T7 arms — the margin arms collapse into negative progress. Right: the exact-counterfactual wall audit — walls are 87% necessary at m=0 but only 49% at m99, against the registered 0.7 bar.

all five held-action alternatives are simulated in the **true** dynamics (others keep their greedy policies). The wall is **necessary** — no alternative survives — in **87%** of the m=0 arm’s walls, but only **49%** of the m99 arm’s (T8-1, FAIL against the registered 0.7). The margin does not merely disable the guard; it **poisons the abstain path’s existence claim**. This is a logical asymmetry worth stating on its own: a clearance margin is a *one-sided* certificate — it can certify negatives (“this action is safe up to ϵ ”) but never positives (“no action is safe”), because the positive claim is made of the very distances the margin has inflated. The companion attack (T8-2) failed to land: the margin premise transports on-policy (m95 ratio **0.83** — drift is *more* predictable, not less), so the theorem keeps its form with the transport measured rather than assumed.

The amendment, and what it restores. The two-tier guard (H-ATK2) implements the asymmetry directly: proposals are certified at the margin (soundness), but the abstain alternative is certified at **true distance** (wall_margin = 0), and the wall fires only on *necessary* failure. The result is a clean separation of what the fix can and cannot do: manufactured walls drop by 46% (1500 → 808) with certified contacts still **0** — the *claim* is repaired and proposal-path soundness is untouched — but progress is **identical** to the single-tier arm (−3.20). The reason is structural, and it closes the loop with §4.5: with an abstain set of one action (STAY), correcting *which states are truly unwinnable*

Table 12: H-ATK2 (tier.json): the two-tier guard removes 692 manufactured walls without touching proposal soundness; capability stays frozen because the abstain set has one element — honesty is not capability.

arm	wall margin	wall events	certified contacts	progress
AR99 (one-tier)	1.867	1500	0	−3.20
AR99t (two-tier)	proposals 1.867 / abstain 0	808	0	−3.20
A0 (reference)	0	483	0	+16.47

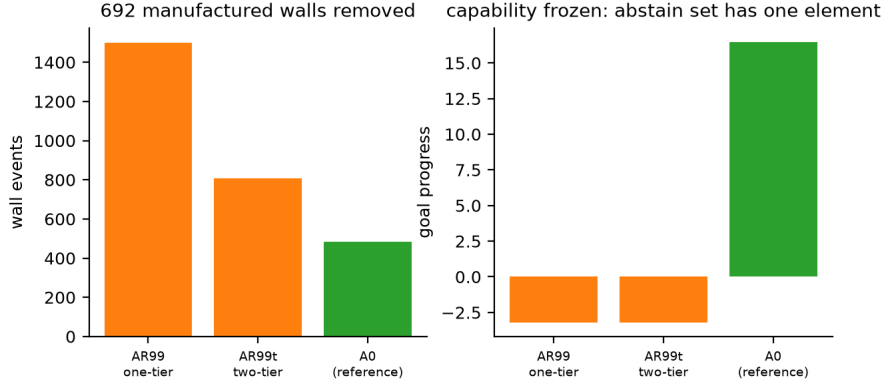


Figure 15: The two-tier guard (H-ATK2): manufactured walls removed (left) while capability stays frozen (right) — with a one-element abstain set, correcting *which states are unwinnable* cannot change *what the system does there*.

cannot change *what the system does there* — both tiers execute STAY either way. **Honesty about unsafety is not capability; capability requires the abstain set to contain maneuvers** — the same lesson highway’s abstain wall taught (§4.5), now derived from the opposite direction: fix the truth-telling first, then grow the set. (Table 12, Figure 15.)

The abstain trilogy: each fix isolates one defect. §4.9’s two-tier guard repaired the wall’s *existence claim* but left capability frozen, because the abstain set held a single action. Three registered attacks grow and then choose it — and each failure names exactly one thing the naive design got wrong (Table 13, Figure 16).

Availability is not enough (H-ATK3). Growing the set to {STAY, evade} — where evade steps away from the nearest opponent — removes 71% of the manufactured walls (808 → **232**, AT3-1 PASS) with certified contacts still **0** (AT3-2 PASS), but progress collapses to −47.92 (AT3-3 FAIL, spectacularly). Greedy evasion is **exploitable**: fleeing the nearest opponent herds the ego into the reach of the next, so the abstain maneuver must be chosen against a *reacting* opponent, not a static one.

Game-theoretic choice: the walls vanish, soundness breaks (H-ATK4). Choosing the abstain maneuver by predicted worst-case clearance under the chase simulation (max–min) drives walls to **0** — vacuously, since every state now finds a predicted-safe maneuver — but certified contacts rise to **4** (AT4-2 FAIL) and progress to −88.02 (AT4-3 FAIL). The max–min maneuver maximizes *flight*, not *task*: the ego retreats, and the maneuver that maximizes predicted clearance is executed unsound at the error’s tail. This isolates the **objective defect** — an abstain maneuver must be chosen for the task.

The doctrine, and the margin’s magnitude (H-ATK5). The three-way doctrine is now implemented literally: **existence** at true distance (the wall is honest), **execution** certified *at the margin* (soundness), and — among margin-feasible candidates — the one minimizing predicted distance-to-goal (**task-first**). The result: walls **1500**, certified contacts **0**, progress −3.20 (AT5-1/2/3 FAIL). The failure is not a mechanism but a **magnitude**: at m99 = 1.867 *no* abstain maneuver certifies at all (0 of 1500 steps), so the guard falls to the wall every step and executes STAY — identical to the single-tier arm. The catch-22 is now fully visible: execution at true distance is **unsound** (AT4’s 4 breaches), execution at m99 is **infeasible** (AT5’s 1500 walls), and the wall’s existence claim requires true distance. A

Table 13: The abstain trilogy (grown.json, atk4.json, atk5.json): availability (H-ATK3) fixes walls but is exploitable; the game-theoretic choice (H-ATK4) removes walls but breaks soundness and the objective; the full doctrine (H-ATK5) is *correct* but the scalar margin at m99 makes execution infeasible — the failure migrates from mechanism to magnitude.

attack	abstain design	walls	certified contacts	progress
H-ATK2	{STAY}, two-tier	808	0	−3.20
H-ATK3	{STAY, evade}, greedy	232	0	−47.92
H-ATK4	{STAY, evade}, max-min clearance	0	4	−88.02
H-ATK5	{STAY, evade}, margin-certified, task-first	1500	0	−3.20
A0 (reference)	—	483	0	+16.47

Figure 16 — the abstain trilogy: each fix isolates one defect (availability → soundness → objective → feasibility)

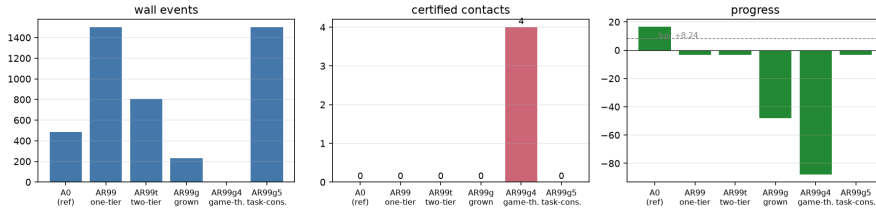


Figure 16: The abstain trilogy (Table 13): walls fall as the abstain set grows and is chosen game-theoretically (left), but certified contacts reappear at H-ATK4 (centre) and progress collapses (right) — each arm isolates one defect. H-ATK5’s return to 1500 walls is the margin’s magnitude, not a design error: at m99 no maneuver certifies.

scalar margin cannot be both $\geq \varepsilon$ (soundness) and $\ll$ the interaction distance (feasibility), because ε ’s tail (1.9 world units) dwarfs the contact radius (0.3). It is feasible only if ε itself shrinks — the **structure-aware (tube) margin** §4.8 derived from the dynamics side, now forced from the abstain side. Two substrates, two error sources, one requirement.

The circularity closes: intent without privilege. The responder model of §4.9 used the others’ *known* policy — a privileged tag the bias–predicate theorem cannot assume. T8-3 removes it. A seeded logistic classifier reads each other’s intent (ego-directed chase vs landmark-directed) from a **W = 10 relative-motion history** alone — features ego-directedness, goal-directedness, closing speed, bearing rate, speed, and distance — with held-out accuracy **0.872** (T8-3a PASS, bar 0.85; 2583 train / 861 test). Substituted into the responder model, the inferred ε m95 is **0.751** against the privileged **0.905** and CV **1.662** — a recovery of **1.20×** of the privileged gain (T8-3b PASS, bar 0.5). **Privilege is not required:** intent is inferable from motion alone (Table 14, Figure 17). *Disclosure:* recovery > 1 means the inferred model **beat** the privileged one — the oracle tag is not the optimum of the response family, so the correct reference bound is *the best response predictor in the family*, not “privileged intent.”

Is the safety spec regime-relative? Yes — and stricter is not safer. T8-4 crosses three predicates — **P1** contact-free (violation iff distance < 0), **P2** separation (violation iff raw distance < 0.2), **P3** bounded-contact-time (violation iff contact persists > 5 consecutive steps) — with the three scenarios, arm A0* (execute the deployment action if it is certified safe under the predicate within the horizon, else STAY). The best predicate by (fewest violations, then most progress) **differs across scenarios** — P3, P3, P1 (Table 15, Figure 18) — so **safety is regime-relative** (T8-4 PASS, ≥ 2 distinct optima). Two disclosures. First, in tag *no* predicate is violated (the ego never contacts), so that scenario is vacuous for the spec question — its “P1 optimum” is a progress tie-break among three zero-violation predicates. Second, and counterintuitive: **P2 (separation) is dominated everywhere**, and in adversary it has *more* raw violations than P1 (**476** vs **366**). Tightening the contact predicate does **not** buy safety, because the larger forbidden zone forces more abstains, and the abstain maneuvers are themselves contact-producing. **A stricter safety predicate can increase raw violations** — the monotonicity intuition (stricter $\Rightarrow$ safer) is false in this regime.

Table 14: T8-3 (t83.json): the intent signal survives the removal of privilege — a motion-only classifier recovers (and slightly exceeds) the oracle-tagged model’s error reduction.

responder model	intent source	ε m95
constant velocity	none (rung 1)	1.662
responder (privileged)	oracle policy tag	0.905
responder (inferred)	classifier from motion	0.751

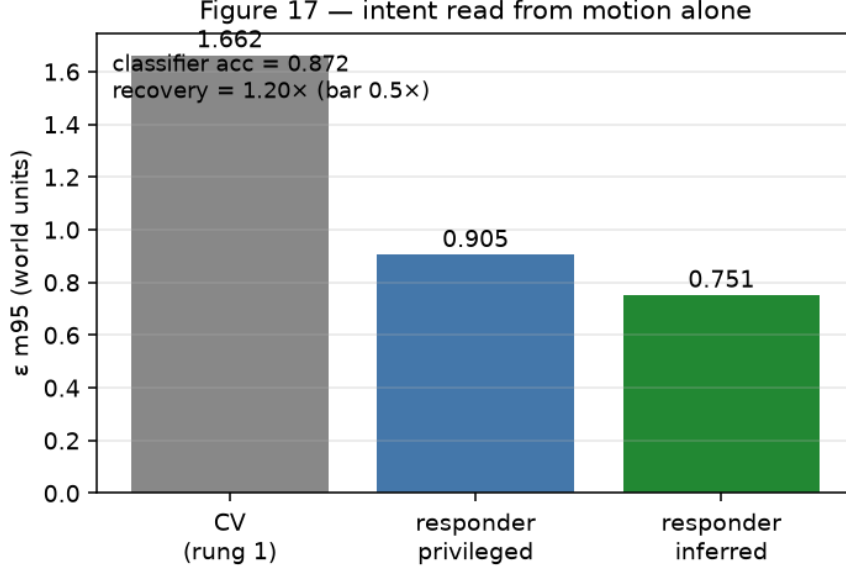


Figure 17: Privilege-free intent (T8-3): ε m95 for the CV, privileged-responder, and inferred-responder models on the same calibration trajectories. A classifier reading a 10-step motion history (accuracy 0.872) more than halves the CV tail and matches the oracle tag — the responder model needs no privileged intent.

Taken together with §4.8, Thm 5’s empirical form now spans three error sources: **dynamics** (highway: exhausted at rung two), **intent** (MPE: bought at $3.4\times$, residual = contact physics), and — named for the next substrate — **perception**. The margin protocol itself survived every attack unchanged; what each attack refined is *where the margin may be used* (proposals) and *where it may not* (existence claims). The abstain path went further and named a *quantity*: it needs the model **three times** (existence, execution, synthesis) and the margin **once per job**, and a scalar margin at the error’s m99 quantile satisfies soundness and feasibility only if the error itself is structure-aware — the requirement §4.8 reached from the dynamics side and §4.9 reaches from the abstain side. That convergence, not any single arm, is the section’s result: the margin’s *form* is a theorem-level obligation, not a tuning choice.

4.10 The monitor: from predicting failure to checking contracts

The last component of Ω — and the one this paper’s first registration left failing — is the runtime monitor. H-3 scored a **pre-execution** scene signal (guard refutation, scene TTC, target-lane occupancy) against the label that the committed maneuver’s postcondition fails within a four-step window. It failed twice: AUROC **0.447** with a score/label deviation the correction ledger preserves (the monitor scored the *proposed* action while the label described the *executed* one), and **0.530** protocol-correct — chance (886 windows, 392 positives). The negative is structural rather than a tuning matter: the surprise that breaks a maneuver has not happened yet at decision time, so no function of the scene at decision time can see it.

Table 15: T8-4 (t84.json): predicate violations / progress per cell (A0*). The best spec differs across scenarios (bold) — regime-relative — and P2 (a stricter contact predicate) is dominated everywhere, with *more* violations than P1 in adversary: enforcement of a wider forbidden zone generates more contact.

scenario	P1 contact-free	P2 separation	P3 bounded-contact-time
adversary	366 / 1.317	476 / 1.181	35 / 1.782
spread	43 / 1.121	90 / 1.769	0 / 1.510
tag	0 / 14.037	0 / 14.037	0 / 13.508

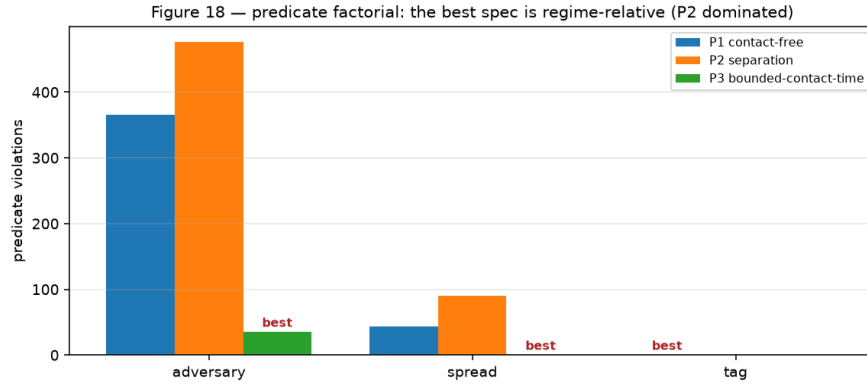


Figure 18: The predicate factorial (T8-4): violations per scenario under the three predicates, with the best cell marked. The optimum is regime-relative (P3 in the two contact-exercising scenarios, P1 — vacuously — in tag), and P2’s wider forbidden zone *raises* raw violations rather than lowering them.

H-MC registers the object the framework’s mapping always described: a maneuver is a **contract** — an intent postcondition (the same claim form §4.5 found the abstain path missing) plus the guard’s safety claim for the executed action — and monitoring is checking that contract against the **partial trace** (addendum H-MC, registered before the run). Everything else is inherited from H-3 unchanged — arms, perturbation injection, window and censoring, replan budget, AUROC estimator — so the comparison isolates the monitored quantity alone. The score at each in-window step is computed *causally from the partial trace*: lateral progress toward a committed lane change, delivered speed change against the committed ± 1 m/s, an IDLE commitment whose lane moves anyway. The pre-outcome rule is fixed in the registration: a window’s score is the maximum over scenes *strictly before* the outcome — neither the crash step nor the postcondition-maturity scene contributes. Detection must be early, or it does not count.

The object change lifts detection from chance to **0.797** (890 windows, 393 positives; registered bar 0.90; Table 16, Figure 20) — a +0.267 AUROC gain from the same trials design, and the residual is identifiable rather than diffuse. Windows opened under a *satisfied* contract and then struck by the injected surprise — the mid-window hard brake or cut-in — carry no contract-violation evidence before maturing, because the surprise is not a breach of any clause the contract contained. **Contract monitoring checks what was contracted; it cannot foresee the exogenous.** The recovery arm (MC-2) misses its bar at exactly H-3’s rates — 28.6% against a 35.7% baseline, −7.1pp against the −5pp bar — and the redundancy is diagnostic: 50 replans against H-3’s 16 produce *identical* success rates, so recovery’s binding constraint is the replan’s efficacy within budget 3, not the trigger that calls it. The component is now bounded from both sides: chance for scene-level prediction, 0.797 for trace-level contract checking, 0.90 still open — and the bound’s *shape*, contract-relevant failure detectable and exogenous surprise not, is the monitor’s characterized frontier rather than an unmeasured guess.

Table 16: The monitor’s object, ablated (monitor_asdev.json, monitor.json, contract_monitor.json). Moving the signal from the scene at decision time to the guard’s contracts checked against the partial trace is the entire difference between chance and near-detection; recovery is invariant across all three rows, isolating its constraint outside the monitor.

monitor	signal (object)	AUROC	windows / positives	recovery baseline	vs	replans
H-3 as-deviated	pre-execution scene	0.447	—	−10.7pp		65
H-3 protocol-corrected	pre-execution scene	0.530	886 / 392	−7.1pp		16
H-MC	contract vs partial trace	0.797	890 / 393	−7.1pp		50
registered bar	—	≥ 0.90	—	$\geq -5\text{pp}$		$\leq 3/\text{ep}$

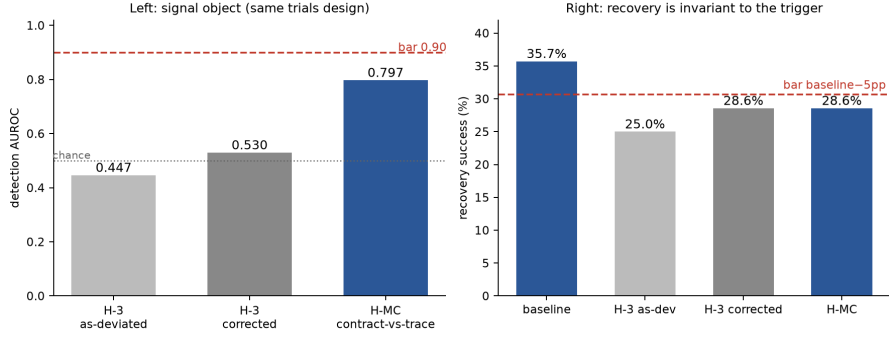


Figure 20: The monitor’s object, ablated (H-3 vs H-MC). Left: detection AUROC — the pre-execution scene score sits at chance (0.447 as-deviated, 0.530 corrected) against the contract-vs-partial-trace check at 0.797, with the registered 0.90 bar. Right: recovery success — baseline 35.7% and both monitor arms 28.6%, identical to one another; the −5pp bar fails in both, and 50 replans buy nothing over 16.

5 The law

Ten results, one law: **the safety boundary is the authority’s model, not the proposer**. The one-step guard’s failure was a defect in the model’s *horizon* (§4.1); the lookahead guard repaired it and exposed the model’s *dynamics* as the residual (§4.2); the corpus’s key had to match the situation, not the scenario (§4.3); and the distilled claims had to be certified against the model’s reachable set, not a sample (§4.4); the abstain path needed the same horizon-soundness as acceptance, and its failure exposed the model’s *abstain set* (§4.5); the measured regime governs a claim’s benefit (§4.6); the ablation splits the layers — safety in the dynamics model, competence in the predicates, speed in the proposer (§4.7); the margin ladder prices the model’s fidelity in capability, rung by rung (§4.8); and the intent rung prices what kinematics cannot — where the attack surfaced the one-sidedness of margin certificates (§4.9), the abstain trilogy isolated the margin’s *form* as a theorem-level obligation, privilege-free intent removed the responder model’s oracle tag, and the predicate factorial showed the *spec* itself is regime-relative (§4.9); and the monitor changed *object* — from a pre-execution prediction of failure (0.530, chance) to a check of the guard’s own contracts against the partial trace (0.797), which bounds it rather than perfects it (§4.10). Each falsifier named a specific piece of the model to strengthen — and the strengthened piece is always the authority.

This is the same law the theory paper states for learned heuristics (*choose the bias for the predicate*), now observed on the safety side: the model that the guarantee rests on is the thing that must be right, and improving it is what the framework buys.

Table 17 is the law as a journey: one row per turn of the loop §3 stated, the registered falsifier each turn, the component of Ω it named, the price measured after repair, and the section that reports it. The nine rows are exhaustive over §4 — and the tenth, the registered perception substrate, is reported as obligation, not evidence.

Table 17: The journey: ten turns of the loop (§3), one row per §4 falsifier. Every row’s bar was registered before its run; every price is read from the artifacts listed in §8; the last row is the registered next substrate — reported as obligation, not evidence.

turn	falsifier (registered before run)	component of Ω strengthened	measured price	§
attack	GS-1, held-out proposers	guard’s horizon ($k = 1 \rightarrow k = 3$)	27 $\rightarrow$ 2 ego-initiated (93%)	4.1–4.2
attack	CT-1, certificate semantics	claims certified against the reachable set, not a sample	748/5369 $\rightarrow$ 0 violations	4.3–4.4
attack	abstain-path audit (H-CT/H-S)	fallback certification	93% of crashes on unchecked fallbacks $\rightarrow$ certified-or-disclosed	4.5
measure	H-RP, dose–response; H-6-D	regime prior R , recorded not predicted	$\rho = 0.79$	4.6–4.7
attack	H-T5, H-T6, model ladder	margin = $\hat{\varepsilon}$ at the kinematics rung	CV $\equiv$ CA (ratio 1.0008): budget exhausted	4.8
attack	T7; H-ATK2–5, abstain trilogy	responder model; three-way abstain doctrine	intent buys 3.4 $\times$; scalar margin named a theorem obligation	4.9
measure	T8-3, privilege-free intent	intent without the oracle tag	acc 0.872; recovery 1.20 $\times$	4.9
attack	T8-4, predicate factorial	the spec itself	regime-relative; stricter $\neq$ safer	4.9
attack	H-3 $\rightarrow$ H-MC, contract vs trace	the monitor’s object: scene $\rightarrow$ partial trace	AUROC 0.530 $\rightarrow$ 0.797 (bar 0.90 open); recovery invariant	4.10
next	perception channel (registered)	model M’s third error source	—	6

The law needs its regime qualifier. The bias experiment that missed its crossing bar (matching-bias 1.02 $\times$ classical, reported FAIL) is explained by measurement, not rescued by caveat: per scenario, the classical–matching expansion gap tracks 1 – FASTER-admissibility at $\rho = +0.79$. Where the regime makes a predicate-blind bias optimal (roundabout, 97% admissible: matching costs 1.5 expansions), matching cannot win; where it does not (highway, 36%: matching wins +0.63), it wins decisively. The aggregate that failed (1.02 $\times$) averaged opposing regimes — the regime lives in the scenario’s traffic structure, not in density. This is Theorem 4’s triple $(\mathcal{H}, P, R)$ observed on the safety substrate: **without R the law is falsified by its own aggregate; with R , every cell falls on the line.**

The law’s last corollary — the fidelity budget. §4.8 adds the economic form: **capability $\propto$ model fidelity, priced by a margin equal to the measured error** — and in interactive traffic that fidelity is bounded by kinematics, exhausted at rung two (CV $\equiv$ CA, ratio 1.0008). What the guard cannot buy at any kinematic order is the *reaction* of the others; the residual budget of containment is **intent**. The law and its corollary now name the same destination from both sides: the model is the deliverable, and the next model is not a better extrapolator — it is a responder. §4.9 supplies the two properties that make the responder a *framework* object rather than a substrate artifact: its intent is **inferable without privilege** (T8-3), and the spec it enforces is **regime-relative** (T8-4) — while the margin it consumes must be **structure-aware** to be both sound and feasible (§4.9).

6 Limitations and open problems

The approximate-authority residual — characterized. The certified classes (2 proposal- plus 5 abstain-certified) were priced in §4.8: a clearance margin equal to the model’s measured horizon error eliminates them, but on this substrate the scalar that covers the constant-velocity model (m95 = 12.9 m) disables the guard, and the constant-acceleration ladder rung reduces the error by a ratio of 1.0008 — kinematics is exhausted. What remains is no longer an open question but a *characterized* frontier: containment under an approximate authority holds at margin = measured error, capability tracks fidelity, and the fidelity budget in interactive traffic is spent on intent, not kinematics. §4.9 climbed the intent rung and sharpened the residual further: on MPE the responder model buys **3.4 $\times$** on the mean error, its intent is recoverable without privilege (T8-3), and its containment residual is

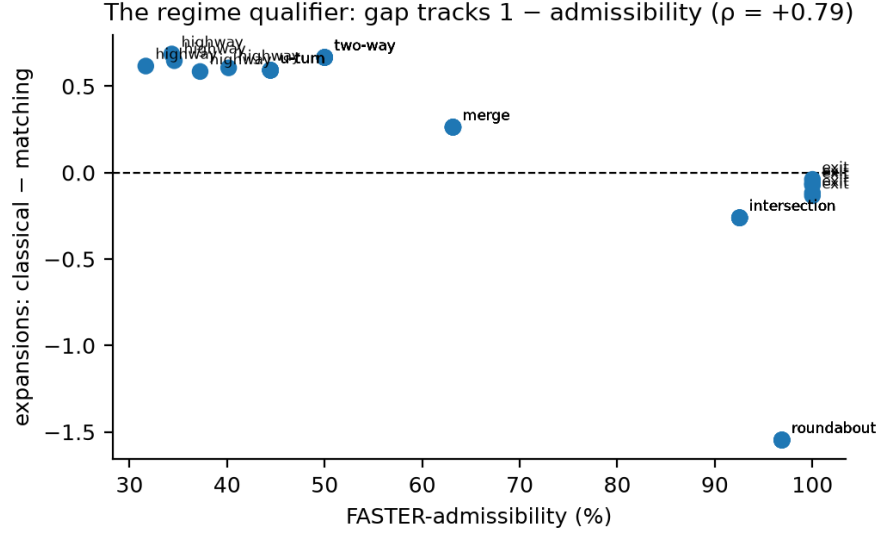


Figure 9: The regime qualifier (H-6-D): the classical–matching expansion gap against FASTER-admissibility, per scenario cell ($\rho = +0.79$). Left of the midpoint the matching bias wins; right of it, classical’s blind speed-first order is optimal and matching is pure overhead.

contact physics — but the abstain path then exposed a *quantitative* requirement the scalar margin cannot meet (AT4/AT5): the margin must be $\geq \varepsilon$ to be sound and $\ll$ the interaction distance to be feasible, which a scalar at ε ’s m99 quantile (1.9 world units against a 0.3 contact radius) cannot be. **The remaining work is now precisely named: a structure-aware (tube) margin whose width is a function of the error’s structure, not its quantile** — reached independently from the dynamics side (§4.8) and the abstain side (§4.9). The theorem’s remaining task is the intent-tube model (a *responder*, not an extrapolator) and its tube-margin proof — and the substrate where others are agents, i.e. the multiagent probe and, for the perceptual error source, the next one.

The abstain set — closed as a mechanism question. Horizon-certifying the fallback changed accounting, not outcomes: within {SLOWER, IDLE} the wall fires about once per episode (11 wall-class crashes / 28). §4.9 grew the set and chose it game-theoretically, and the trilogy *closed* the mechanism question: availability (H-ATK3), soundness (H-ATK4), and objective (H-ATK5) each isolate one defect, and the full three-way doctrine is *correct* — its residual failure is the scalar margin’s magnitude, not the abstain set’s design. The open item therefore migrates from “grow the set” to the **structure-aware margin** named above: with a tube margin the margin-feasible abstain set becomes non-empty, and the H-ATK5 wall disappears without a further mechanism.

The monitor’s remaining limits. Both registrations are reported in §4.10: as a *predictor* the monitor is at chance (0.530); as a *checker* of the guard’s contracts against the partial trace it reaches 0.797 — bounded, not solved. What stays open is precisely what the bound isolates. The exogenous surprise falls under no clause the ego’s contracts contain, so detecting it would require contracts over the *environment*’s commitments (right-of-way, braking distance, lane discipline) rather than the ego’s — a different registration again. And recovery remains gated by the replan’s efficacy, not the trigger: within budget 3, the certified search does not rescue perturbed episodes regardless of when it is called.

The learned proposer’s features. Capacity was not the binding constraint: $4\times$ data moved the linear model +2.6pp and a 64-unit MLP ceilinged at 0.745 agreement (bar 0.80). The fingerprint’s aliasing is — richer features (explicit TTC/gap/leader structure) are the next move (H-PQ3), not bigger models.

Closed: the regime prior. §4.4’s named next move is measured and passed ($\rho = 0.79$, target-measured beats mismatched, containment held). Certification says the claim is true; the measured regime says where it matters — and the measurement is an environment fact the authority records.

The attribution heuristic. Ego- versus other-initiated is decided by a documented geometric heuristic (was the contact partner ahead of the moving ego’s heading). It is a reasonable proxy, not ground truth, and a finer attribution would sharpen §4.1.

Scope. Simulation only; sim2real is out of scope. The continuous scenarios have no discrete guard and are recorded as such rather than silently included. The monitor’s contract semantics and the grown abstain set are registered but unrun.

7 Conclusion

We attacked a structural guarantee on a real substrate rather than assume it, and the attack succeeded twice: a one-step guard is not sound containment for dynamics with memory (27 ego-initiated against a bar of 0), and — once every executed action was classified — the repaired guard’s *abstain path* was not sound either (93% of the remaining crashes lived on unchecked fallbacks). Each repair is the same shape: extend the horizon-soundness requirement to what the guard actually executes, accepted or abstained. What survives both repairs is the model’s own error — and §4.8 priced it: a margin equal to the model’s measured error eliminates the certified classes but, on this substrate, buys containment with the brake pedal; the ladder rung that was supposed to buy it back (constant acceleration) reduces the error by a ratio of 1.0008. The fidelity budget in interactive traffic is spent on kinematics by rung two. What remains is intent — which is why the programme’s next substrate is the one where others are agents, and why the measuring instrument built here (measure the model, certify at its error, report the walls) travels with it.

The efficiency side obeys the same law: the key must be structural, “certified” must mean certified against the reachable set, and a claim’s *benefit* is governed by the substrate’s measured visitation ($\rho = 0.79$) — an environment fact, recorded, never predicted. The ablation says what each layer is for: safety lives in the dynamics model (invariant to predicate corruption), competence in the predicates, speed in the proposer — where a cloned MLP eventually earned its seat ($20.16 \geq 19.81$ m/s) while the certified safety classes stayed literally identical across every proposer quality. What remains is not more data but a better model — of the dynamics, of the abstain set, and of the trace a maneuver’s contract leaves behind.

References

- Eduardo P. Denadai. Exact-authority belief-space planning: Where composition emerges and where it fails, 2026a. Optimality.ai.
- Eduardo P. Denadai. Learnable boundaries, unlearnable competence: A characterization of learned components in a symbolic search stack, 2026b. Optimality.ai.
- Eduardo P. Denadai. Containment rather than calibration: A sound architecture for learning under partial observability, 2026c. Optimality.ai.
- Eduardo P. Denadai. The cost of soundness: f-contour bounds and the inductive-bias limits of learned heuristics, 2026d. Optimality.ai.

8 Provenance

Every number and figure is read from `env_substrates/highway/results/{scenarios, safety, transfer, fragments, fragments_library, handshake, grounding, bias, monitor, contract_monitor, regime, surgical, proposer, bias_regime, abstain, capacity, thm5, thm6}.json`, and for §4.9 the multiagent substrate’s `{h0_adversary, epsilon, ladder, attack, tier, grown, atk4, atk5, t83, t84}.json` (rendered offline by its own `render_mpe.py` — a pure-JSON viewer with the EGO marked) by `make_figures.py`; the substrate code is in `env_substrates/highway/` (`authority.py`, `highway_env_adapter.py`, `corpus.py`, `fragments.py`, and the `run_*.py` drivers). Bars were registered in `env_substrates/highway/docs/PLAN.md` before each run — including the closure addenda (H-ORE, H-RP, H-SC, H-PQ, H-6-D, H-S3, H-PQ2), the Thm-5 arc (H-T5, H-T6), and the contract monitor (H-MC, registered before `run_contract_monitor.py`);

§4.9's bars (M-EP, T7, T8-1/2, H-ATK2-H-ATK5, T8-3, T8-4) were registered in `env_substrates/multiagent/PLAN.md` before their runs, with the T8-3 recovery-ratio miscalibration and the T8-4 degeneracy disclosed in the text; failed bars are reported where they occurred (Tables 2, 3, 4, 6, 7, 8, 9, 10; H-4's VOID, H-6's and PQ-1's and PQ-2-0's and T5-2's and T6-1's and T6-3's FAILs). The correction ledger (the tau-wrapper defect and its preserved `*_astau0.json` as-run artifacts; the monitor score/label deviation; PQ-2's registration ambiguity; the H-T5 id-matching fix and the disclosed cross-process ϵ drift) lives in the substrate `PLAN` and §4.2. The paper's own section/figure set was registered in `PLAN.md` before these sections were written. Figures: `fig1` (where containment breaks), `fig2` (guard activity), `fig3` (corpus growth), `fig4` (transfer), `fig5` (retention), `fig6` (one-step vs lookahead), `fig7` (abstain taxonomy), `fig8` (regime dose-response), `fig9` (regime qualifier), `fig10` (proposer seat), `fig11` (margin economics), `fig12` (model ladder), `fig13` (intent ladder), `fig14` (capability and wall), `fig15` (two-tier guard), `fig16` (abstain trilogy), `fig17` (privilege-free intent), `fig18` (predicate factorial); `fig16`–`fig18` are generated by `imgs/make_figs_49b.py` from the multiagent artifacts, and `fig20` (the monitor's signal object and recovery, from `monitor{,_asdev,contract_monitor}.json`), and `fig19` (the Ω -TD architecture — a **structural** diagram carrying no data, registered in `PLAN.md` addendum ARCH before generation, source `fig19.d2`, rendered by the D2 compiler and invoked by `make_figures.py`). The paper's architecture frame (§3: the two invariants, the loop), §4.10 and its Table 16, and the journey table (Table 17, §5 — ten registered turns, one per §4 falsifier, placed in §5 to keep the table sequence 1–16 monotone; recorded as a placement correction in `PLAN.md`) are registered in the same addendum. Contract tests: `env_substrates/highway/test_highway.py` (9 checks) and `env_substrates/multiagent/test_multiagent.py` (3 checks) — all passing. The experiment trace lives in the `PLAN` and the artifacts, not here.