RISK BOUNDS FOR PHYSICAL AI: DECISION AMBIGUITY AND MODEL INCOMPLETENESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical AI safety depends not only on what a system perceives, but on what its action will do and whether a response can still work in time. From classical decision, identification, and testing results we compose one risk bound for a single irreversible decision that separately budgets the loss of continuing while the consequence model is valid, of continuing after it has silently failed, and of responding, including responses that fail or come too late. A hand-constructed finite example computes every term of the bound, and a synthetic model instantiates its monitoring term. A MetaDrive study, which does not estimate the bound, asks whether detect-and-brake monitors pass a preset intervention screen in intact driving and how response choice changes the balance between repaired and newly induced failures. A hidden fault that scales braking commands to 70% produces 21 failures in 120 drives selected for nominal success. None of twelve detect-and-brake candidates frozen before the audit passes the screen: the least-triggering ones intervene in 3 of 24 intact scenes where at most 2 were allowed; with 24 scenes this fails the screen but cannot show that the true rate exceeds 10%. Offline re-scoring with a post-hoc predictor whose one-step error is 37.7% lower detects more faults, but its best variants still cross the threshold in 3 of 24 intact scenes. In post-hoc same-scene comparisons, compensation with the true fault gain, started at the known onset in fault drives only, repairs 15 of 21 baseline failures and introduces two; started by an onset-aligned trigger in every drive, it leaves 16 fault failures and causes 14 intact-drive failures. Three braking responses on that trigger each introduce more failures than they repair in point estimates (paired intervals include or touch zero). Together, the analysis and experiments motivate treating information sufficiency, consequence validity, and response execution as separate assurance obligations.

1 INTRODUCTION: WHY A CORRECT OBSERVATION IS NOT ENOUGH

Physical AI systems, learned controllers that steer a car, lift a patient, or deliver a package, take actions that cannot be taken back. A care robot that misjudges its own grip can drop an elderly person and break a hip. A delivery robot whose motors have worn, or whose controls have been tampered with, can stall in rush-hour traffic or roll into it. A vehicle can identify every pedestrian and obstacle correctly and still crash because its brakes have weakened while its controller continues to act as before. In each example, the system relies on a prediction of what its actions will do. The calculation that declared the action safe can be correct while its premise is wrong. Better perception of the road does not, by itself, establish the vehicle's current braking power.

Why an alarm is only part of the answer. One natural response is to detect a discrepancy and switch to a conservative action. This helps only if the warning arrives while a useful response remains available, and the response has its own consequences: stopping in traffic can create a new hazard, and unnecessary interventions can block the task. The practical question is therefore not just "Did the detector notice a fault?" It is *"Did the complete mechanism reduce danger without introducing unacceptable harm or disruption?"*

Three questions for a safety argument. **Information:** does the system know enough to choose an acceptable action? **Prediction:** what evidence supports its estimate of the action's consequences? **Execution:** can the chosen response still take effect in time? We call these information sufficiency,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

consequence validity, and timely execution. Two failure sources run through the paper: *decision ambiguity*, when no single action is acceptable in every state the evidence allows, and *model incompleteness*, when the system's model of what its actions will do silently stops being accurate. A remedy for one need not satisfy the others: a clearer camera view does not restore worn brakes.

What the experiments reveal. We use MetaDrive, a driving simulator with generated roads and traffic (Li et al., 2023), and an unchanged driving policy released by other authors (Cai et al., 2025). It was trained using proximal policy optimization (PPO), a reinforcement-learning method (Schulman et al., 2017). Scaling its braking commands to 70% produces 21 failures across 120 drives selected for nominal success; the effect repeats in two disjoint cohorts. None of twelve detect-and-brake candidates frozen before the audit passes its screen: the least-triggering ones intervene in 3 of 24 intact scenes where at most 2 are allowed; with 24 scenes this fails the screen but cannot show that the true rate exceeds 10%. Offline re-scoring with a post-hoc predictor whose one-step error is 37.7% lower detects more faults, but its best variants still cross the threshold in 3 of 24 intact scenes. In post-hoc comparisons on the same scenes, three braking responses on a trigger aligned with the fault onset by construction end with 26–32 fault failures, against 21 with no response. Command compensation with the true fault gain ends with 8 when started at the known onset in fault drives only, and with 16, plus 14 intact-drive failures, when started by that trigger in every drive. No tested detector passed the screen, and even with trigger timing supplied, *fault failures fell under compensation and rose under braking* in point estimates.

A running example: one tenth of a second matters. Consider a vehicle approaching a stationary obstacle. It travels at speed v, has distance d remaining, and waits L seconds before braking takes effect. If evidence shows that braking achieves at least deceleration $\underline{b} > 0$, the stopping condition is (cf. Shalev-Shwartz et al., 2017)

$$vL + \frac{v^2}{2\underline{b}} \leq d. \tag{1}$$

The first term is distance traveled while waiting; the second is braking distance. At $v = 10\,\mathrm{m/s}$, $d = 25\,\mathrm{m}$, $L = 0.5\,\mathrm{s}$, and $\underline{b} = 2.5\,\mathrm{m/s^2}$, stopping needs exactly $25\,\mathrm{m}$. Another $0.1\,\mathrm{s}$ raises this to $26\,\mathrm{m}$. The same questions arise for care and delivery robots; we test only driving.

Contributions. First, one risk bound for an irreversible decision (*commitment*) that adds up, over the same horizon, the loss from insufficient information, from a silently invalid consequence model, and from a response that fails or comes too late (Theorem 5.2). Its constituent Bayes-risk, partial-identification, and total-variation facts are classical; the contribution is composing them under shared assumptions about the consequence model, the available information, and the response. Second, an admission procedure that runs each complete mechanism, detector and response together, in closed loop, screens it on nominal intervention burden and task progress, and treats an empty set of admissible mechanisms as a legitimate outcome (Section 5, with the audit in Section 5.1). Third, a frozen learned-controller stress study followed by explicitly post-hoc response comparisons (Sections 6.2–6.4). It does not evaluate the bound; it checks whether a lower-error predictor reduces nominal triggers enough to pass the screen, and how response choice changes recovery and collateral harm.

2 Choosing an action under a valid model

An occluded crosswalk creates uncertainty about *which action to choose*; weakened brakes create uncertainty about *what an action will do*. We separate them under a model with a justified prediction-error bound. The intermediate Bayes-risk, identification, and total-variation (TV) testing results are classical; we state them in the form Theorem 5.2 needs.

Fix a task, a decision horizon T, and an action family $\mathcal{A}$. An action, or *commitment*, may be a complete short maneuver, including feedback, whose consequences cannot be recalled once started. A loss $\ell(\omega, a) \in [0, 1]$ measures its consequence in outcome ω; for a failure indicator, expected loss is failure probability. The history H_t is everything the controller has observed by time t. The policy may use this history and independent randomization, not hidden state.

For a true environment law P and an available history h of positive probability, compare the actual conditional risk L_P with the model's predicted risk $L_{\mathcal{M}}$:

$$L_P(a \mid h) = \mathbb{E}_P[\ell(\omega, a) \mid H_t = h], \qquad L_{\mathcal{M}}(a \mid h) = \sum_{s \in \mathcal{S}} b_h(s)\ell_h(s, a). \tag{2}$$

The model lists possible states $\mathcal{S}$, their probabilities $b_h(s)$, and their action losses $\ell_h(s, a)$. Model and evaluator use the same task, actions, and loss.

How accurate must the prediction be? A consequence contract states how far predicted risk may differ from actual risk, for *every* candidate action. Its tolerance δ is measured in the same loss units.

Definition 2.1 (Decision-validity contract)**.** The model is δ-decision-valid at (P, h) when

$$V^{\delta}(P, h) = \mathbf{1}\left\{\sup_{a \in \mathcal{A}} |L_P(a \mid h) - L_{\mathcal{M}}(a \mid h)| \leq \delta\right\} = 1. \tag{3}$$

Its validity envelope contains the pairs (P, h) satisfying this condition; cases inside it are *covered*. *Model escape*, the model incompleteness studied here, is failure of this contract.

How much risk remains even with that model? Let $\rho_{\mathcal{M}}$ be the smallest risk available to a controller using this history. It measures decision ambiguity, not optimizer quality.

Definition 2.2 (Unavoidable covered-model commitment risk)**.** For finite $\mathcal{S}$ and $\mathcal{A}$, define

$$\rho_{\mathcal{M}}(h) = \min_{q \in \Delta(\mathcal{A})} \sum_s b_h(s) \sum_a q(a)\ell_h(s, a) = \min_{a \in \mathcal{A}} L_{\mathcal{M}}(a \mid h). \tag{4}$$

For binary loss, $\mathcal{A}^T_{\text{safe}}(s) = \{a : \ell_h(s, a) = 0\}$ is the set of actions safe in state s over the horizon T.

This is conditional Bayes risk; $q \in \Delta(\mathcal{A})$ ranges over randomized actions, although linearity makes one minimizing action sufficient. For example, three equally likely states with safe-action sets $\{a, b\}$, $\{b, c\}$, and $\{a, c\}$ share no safe action although each pair does; their minimum risk is $1/3$.

The next result transfers the model's calculation to the environment. A model-optimal action is within δ of its predicted risk and within 2δ of the true optimum.

Proposition 2.3 (Covered risk and approximate-model transfer)**.** *Suppose Equation* (3) *holds. Every history-based action mixture has true conditional risk at least* $(\rho_{\mathcal{M}}(h) - \delta)_+$*. For a model minimizer* $\hat{a}$*, with a true minimizer* $a^\star$*,*

$$|L_P(\hat{a} \mid h) - \rho_{\mathcal{M}}(h)| \leq \delta, \qquad L_P(\hat{a} \mid h) - L_P(a^\star \mid h) \leq 2\delta. \tag{5}$$

Under exact validity and binary nonnegative loss, a zero-risk action exists if and only if

$$\bigcap_{s : b_h(s) > 0} \mathcal{A}^T_{\text{safe}}(s) \neq \varnothing. \tag{6}$$

Proof: Appendix A.11.

What improves the decision? A better optimizer approaches $\rho_{\mathcal{M}}$; more useful information or a better action family can lower $\rho_{\mathcal{M}}$ itself.

Proposition 2.4 (Value of timely, valid information)**.** *If* $H' = (H, X)$ *refines history under the same exact joint model, action family, and loss, then*

$$\mathbb{E}[\rho_{\mathcal{M}}(H') \mid H] \leq \rho_{\mathcal{M}}(H). \tag{7}$$

Proof: ignoring X remains feasible (Appendix A.11). Waiting for information helps only while acting on it remains feasible, as in the occluded-crosswalk model of Section 6.1.

3 What risk can the available evidence establish?

Before braking, intact and weakened brakes can produce the same observations. If allowed environments remain observationally indistinguishable yet imply different risks, the safety argument must retain those possibilities. This is *partial identification* (Manski, 2003; Pearl, 2013).

Fix a policy π and an information-collection protocol $\mathcal{I}$ through a declared deadline. Let $Q^{\mathcal{I}}_P$ be the distribution of the full available history under $\mathcal{I}$, and $R_P(\pi)$ the policy risk in environment

P. Engineering assumptions define a *completion class* $\mathfrak{P}$: the environments the argument allows. Group environments with identical $Q_P^{\mathcal{I}}$ into a cell c, and collect their possible risks:

$$\begin{aligned} \Gamma_\pi(c) &= \{R_P(\pi) : P \in c\}, \qquad \bar{R}_\pi(c) = \sup \Gamma_\pi(c), \\ \operatorname{diam}_\pi(c) &= \sup \Gamma_\pi(c) - \inf \Gamma_\pi(c). \end{aligned} \tag{8}$$

diam_π is the width of the uncertainty; $\bar{R}_\pi$ is its worst allowed risk. If the possible risks are 0.1 and 0.7, a single report is wrong by at least 0.3 somewhere, and a valid upper guarantee is at least 0.7.

Proposition 3.1 (Contract-relative risk-reporting limit)**.** *For any scalar risk report g based on the shared observable law,*

$$\sup_{P \in c} |g - R_P(\pi)| \geq \frac{\operatorname{diam}_\pi(c)}{2}. \tag{9}$$

The midpoint of the infimum and supremum attains equality. Every uniformly valid upper certificate is at least $\bar{R}_\pi(c)$, which itself is valid. If $g(H) \in [0,1]$ is a history-based report, the same lower bound applies to its worst-case marginal bias $|\mathbb{E}_P g(H) - R_P(\pi)|$ *within c.*

Proof: Appendix A.11. Knowing risk precisely differs from knowing it is low: the singleton $\{0.9\}$ has width zero. Observations help when they distinguish allowed environments; justified restrictions help by excluding some.

Nearly indistinguishable environments. Exact equality has an approximate counterpart. Total variation (TV) measures how differently two distributions can assign probability to the same event; the testing connection is classical (Le Cam & Yang, 2000). If two environments P_0, P_1 have history laws $Q_i = Q_{P_i}^{\mathcal{I}}$ with $\mathrm{TV}(Q_0, Q_1) \leq \varepsilon$ and policy risks that differ by Δ_R, every bounded report $g(H) \in [0,1]$ satisfies

$$\max_{i \in \{0,1\}} |\mathbb{E}_{Q_i} g(H) - R_{P_i}(\pi)| \geq \frac{(\Delta_R - \varepsilon)_+}{2}. \tag{10}$$

This bounds worst-case marginal bias under the fixed information protocol (proof in Appendix A.3). A large risk difference with little observable separation limits what any such report can establish.

An external audit that samples each declared environment improves precision within the class (Appendix A.15); only informative sensing or justified restrictions change which environments remain possible.

4 Will the warning arrive while a response can still work?

The stopping example has two deadlines: when a response must become effective, and the earlier time when the monitor must call for it. Write t_{cert} for the contract's effective-response deadline and $\bar{L}$ for a bound on execution delay. The latest trigger is $t_d = t_{\mathrm{cert}} - \bar{L}$. A conservative contract deadline can precede $t_{\mathrm{phys}}(\omega)$, the latest physical time at which a response could still work. A variable cutoff must be an observation-based stopping rule.

Let Q_0, Q_1 be the benign and hazardous distributions of the full history available through the same fixed or observation-based cutoff. The event D means "trigger by that cutoff." Then $Q_0(D)$ is the false-trigger rate and $Q_1(D^c)$ the hazardous-miss rate. The information available by the deadline limits their achievable tradeoff. The TV-testing identity is classical (Le Cam & Yang, 2000); here it supplies the deadline-limited information term used by the composed bound.

Proposition 4.1 (Deadline-limited monitorability frontier)**.** *For every such event D,*

$$Q_0(D) + Q_1(D^c) \geq 1 - \mathrm{TV}(Q_0, Q_1). \tag{11}$$

The minimum is attained by a likelihood-ratio decision. More generally, for costs $w_0, w_1 > 0$ and densities q_i with respect to $\mu = Q_0 + Q_1$,

$$\inf_D \{w_0 Q_0(D) + w_1 Q_1(D^c)\} = \int \min(w_0 q_0, w_1 q_1)\, d\mu. \tag{12}$$

Proof: Appendix A.11. When the histories are indistinguishable ($\mathrm{TV} = 0$), lowering false triggers forces more misses; better separation permits, but does not guarantee, a better monitor.

Corollary 4.2 (Selective mitigation requires decision-time separation)**.** *A monitor satisfying $Q_0(D) \le \alpha$ and $Q_1(D^c) \le \bar{m}$ requires*

$$\mathrm{TV}(Q_0, Q_1) \ge 1 - \alpha - \bar{m}. \tag{13}$$

If failing to trigger by t_d causes conditional hazardous loss at least ϕ and loss is nonnegative otherwise, then the expected loss R_{haz} under the hazardous law satisfies

$$R_{\mathrm{haz}} \ge \phi\,[1 - \alpha - \mathrm{TV}(Q_0, Q_1)]_+. \tag{14}$$

Proof. Rearrange Equation (11); then use the premise $R_{\mathrm{haz}} \ge \phi Q_1(D^c)$ and nonnegativity. □

The design lesson is to measure information *before* execution time is exhausted. A physical-loss claim from a missed warning also needs the premise ϕ, which no detection score supplies.

5 Putting the safety argument together

The three questions become four engineering steps: keep a useful option available, justify what the response can do, account for its execution, and audit the complete mechanism.

Keep a useful option available. Under a valid consequence model, maintain an action that is acceptable across the remaining plausible states. In the running example, the speed bound derived from Equation (1) is

$$v \le -\underline{b}L + \sqrt{(\underline{b}L)^2 + 2\underline{b}d}. \tag{15}$$

Slowing or increasing separation preserves acceptable futures and time for evidence. Barrier and reachability methods implement such envelopes under justified dynamics and disturbance bounds (Ames et al., 2017; Bansal et al., 2017).

State what the response can achieve. A consequence contract records the operating conditions, available actions, allowed loss, and evidence supporting the prediction. Braking needs adequate traction and authority; steering needs lateral clearance. The record tells an engineer what must be checked again when a sensor, actuator, or operating condition changes.

Account for continuing and responding.

Remark 5.1 (Response-risk budget). Let $Y_f \in [0,1]$ be the loss of a response and G the event that it executes validly and in time. If $\mathbb{E}[Y_f \mid G] \le \epsilon_f$ and $\mathbb{P}(G^c) \le \eta$, then $\mathbb{E}Y_f \le \epsilon_f + (1-\epsilon_f)\eta$. The composed bound applies it only where a response is invoked (standalone form: Appendix A.12).

The whole mechanism has four possible outcomes. The controller either continues or invokes a response. Continuation can occur inside or outside the model's validity envelope; an invoked response can satisfy or fail its validity and timing contract. We bound the loss across all four cases on the same horizon. Fix a decision cutoff $t_d = t_{\mathrm{cert}} - \bar{L}$ and a common information-collection protocol up to that cutoff. Let $H = H_{t_d}$, let the response be invoked on an event D determined by H, and let C denote the possibly latent covered stratum. Write $e = \mathbb{P}(C^c) \in (0,1)$ for the escape mass, Q_C and Q_E for the laws of H given C and C^c, $\alpha = Q_C(D)$ for the covered trigger rate, and $m = Q_E(D^c)$ for the escaped miss rate. This application of Section 4 uses *covered/escaped* laws, not its benign/hazardous labels: covered cases may have positive risk, and ϕ lower-bounds the loss of an untriggered escaped case.

Theorem 5.2 (Composed risk bound for a physical commitment)**.** *Let $Y_0, Y_f \in [0,1]$ be the losses of the primary continuation $a_0(H)$ and the complete response policy, and let $Y_\pi = \mathbf{1}_{D^c} Y_0 + \mathbf{1}_D Y_f$. Assume the primary action family has an actionwise validity contract* conditional on H, C*, with tolerance $\delta(H)$. Define the optimization gap $\gamma(H) = L_{\mathcal{M}}(a_0 \mid H) - \rho_{\mathcal{M}}(H) \ge 0$ and*

$$\begin{aligned} A_- &= \mathbb{E}[\mathbf{1}_C \mathbf{1}_{D^c} (\rho_{\mathcal{M}} - \delta)_+], \\ A_+ &= \mathbb{E}[\mathbf{1}_C \mathbf{1}_{D^c} \min\{1, \rho_{\mathcal{M}} + \gamma + \delta\}]. \end{aligned} \tag{16}$$

Let G mean that the invoked response is consequence-valid, reachable (its entry conditions hold when invoked), and effective by t_{cert}. For $\epsilon_f, \phi \in [0,1]$, suppose $\mathbb{E}[Y_f \mid D \cap G] \le \epsilon_f$ and $\mathbb{E}[Y_0 \mid$

$C^c \cap D^c] \geq \phi$ *whenever the conditioning events have positive probability. With* $\xi = \mathbb{P}(D \cap G^c)$ *and* $R_\pi = \mathbb{E}Y_\pi$,

$$\boxed{A_- + e\phi[1 - \alpha - \mathrm{TV}(Q_C, Q_E)]_+ \;\leq\; R_\pi \;\leq\; A_+ + em + \epsilon_f \mathbb{P}(D) + (1 - \epsilon_f)\xi.} \tag{17}$$

If every failed invoked response is invalid or unreachable (probability at most η_v*) or delayed beyond its budget (probability at most* η_l*), then* $\xi \leq \min\{\mathbb{P}(D), \eta_v + \eta_l\}$.

Proof. Partition the loss over $C \cap D^c$, $C^c \cap D^c$, $D \cap G$, and $D \cap G^c$. Conditional validity and Proposition 2.3 bound the first term between A_- and A_+. The second lies between $e\phi m$ and em. The final two terms have sum at most $\epsilon_f \mathbb{P}(D) + (1 - \epsilon_f)\xi$. Proposition 4.1 gives $m \geq [1 - \alpha - \mathrm{TV}(Q_C, Q_E)]_+$; substitution gives the lower bound. The last claim is a union bound; no independence is required. □

What the composition requires. The proof is a partition plus classical bounds. The parts compose only under three premises: validity conditional on the history *and* the latent covered stratum; one history law and cutoff shared by monitor and continuation; and the premise ϕ, which converts a miss into loss. Conditioning cannot be dropped. Take constant H, $Y \sim \mathrm{Bernoulli}(1/2)$, and $L_{\mathcal{M}} = \mathbb{E}Y = 1/2$, so the unconditional prediction is exact ($\delta = 0$). For the latent event $C = \{Y = 1\}$, multiplying that contract by $\mathbf{1}_C$ would bound covered loss by $\mathbb{E}[L_{\mathcal{M}}\mathbf{1}_C] = 1/4$, but $\mathbb{E}[Y\mathbf{1}_C] = 1/2$ (Appendix A.1).

How to read the bound. A_+ budgets continuing inside the valid model, em escaped cases left without a response, and the final two terms the response itself, including failed execution. Thus better decisions, better timely evidence, and a better response address different terms. Only D, $\rho_{\mathcal{M}}$, and γ are computed online from H. The rates α, m, e, ξ, the bound ϵ_f, and TV bounds are audit quantities, estimable offline from auditable coverage and response labels (Appendix A.13). The tolerance δ, the coverage event C, and the premise ϕ are assurance obligations that need engineering evidence. The bound is an assurance decomposition; an unsupported term is an unmet assurance obligation, not a quantity silently assumed to be zero.

Responses change later observations and therefore later detectability. The stagewise version conditions each stage on the *actually reached* history, including earlier responses; with first-failure increments, the stage bounds sum to an episode bound, with premises at each stage (Appendix A.14).

5.1 Audit the complete implemented policy

A candidate is the whole mechanism: the detector, the response it invokes, and their execution rules. Freeze a finite set Θ before the operational audit. For each candidate, evaluate the full closed-loop implementation, including monitoring duration, trigger rule, queue, latch, and actuation path. Let $I(\theta)$ be nominal intervention frequency, $U(\theta)$ normalized progress (in the campaign, mean route completion over nominal and fault runs), and $R(\theta)$ physical failure risk, with audit estimates $\widehat{I}(\theta)$ and $\widehat{U}(\theta)$. The campaign uses the following *empirical screening rule*:

$$\widehat{\Theta}_{\mathrm{emp}} = \{\theta : \widehat{I}(\theta) \leq 0.10,\ \widehat{U}(\theta) \geq 0.85\}. \tag{18}$$

On 24 scenes, $\widehat{I} \leq 0.10$ allows at most two interventions. The screen checks burden and progress only; declared selectors, one minimizing physical failure, rank eligible candidates, and an empty set returns *no admissible candidate*.

For a population-confidence claim, sample size matters: even zero interventions in 24 scenes has a one-sided exact 95% upper limit of 11.73%, above a 10% budget. Appendix A.7 gives the exact-binomial arithmetic (Clopper & Pearson, 1934) and a simultaneous-candidate rule.

6 Experiments

Table 1 maps each result to the terms of Theorem 5.2 it exercises, separating frozen from post-hoc evidence.

Table 1: What each result tests. MetaDrive rows are frozen before outcomes or post hoc on exposed data; none estimates a term of Theorem 5.2.

Status	Result	Terms	Does not establish
Analytic	Finite example, exact (App. A.10)	All	Any term estimated from data
Synthetic	Decision floors, crosswalk (App. B.1, B.2)	$\rho_{\mathcal{M}}$	A certified validity tolerance δ
Synthetic	48 completions (App. B.4)	α, m, TV	Covered, response terms (zero here)
Synthetic	Mechanism holdout (App. B.6)	m	A physical response
Frozen	Fault stress, two cohorts (§6.2)	None	An effect outside the selected nominal-success, brake-relevant population
Frozen	Audit screen, 12 mechanisms (§6.3)	None	A population-level $\leq 10\%$ intervention guarantee
Post hoc	Contextual predictor, offline (§6.3)	None	Closed-loop admission
Post hoc	Response arms, onset-aligned trigger (§6.4)	None	Fault detection; ϵ_f, ξ with a detector
Post hoc	Diagnostics (App. C.14, C.15)	None	A failure mechanism

6.1 Synthetic models

In a blind-fork model, where the safe side of an obstacle is revealed at a random time, and in a triple-conflict model with three states, simulated decision risk is within .002 and .003 of the analytic floors. In the crosswalk model, preserving a stopping option keeps zero modeled collisions and cuts transit time from 20.5 steps under a fixed conservative speed cap to 13.5 (clear) and 14.5 (fog); the direction is guaranteed because the cap is a feasible policy.

The 48-completion study declares 48 fully specified nominal, benign-novelty, and fault environments (completions). Its endpoint is certification loss, a hazardous environment not flagged by decision step 2 (the latest trigger that stops in time for every modeled braking rate), rather than collision. With non-hazardous cases as the covered stratum (escape mass $e = .20$), the covered and response terms vanish and the bound becomes $.20[1 - \alpha - \mathrm{TV}]_+ \leq R_\pi = .20m$. Rounded inputs give a lower endpoint at least .1526 and risk at most .1529 for a monitor reading one residual channel, and $.1280/.1307$ for an ensemble monitor (Appendix B.4). In a mechanism holdout, a residual monitor designed for a mean-shift fault has miss rate (the experiment's loss) .021 on the matched hazard and .390, .488, and .921 on withheld variance, late, and silent hazards.

6.2 A hidden actuator fault in a frozen controller

We evaluate the unchanged PPO policy released with the AIM interactive-learning method (Cai et al., 2025) in MetaDrive (Li et al., 2023). At the onset t_e of the first sustained braking in the intact run (the first decision index ≥ 20 that begins three consecutive requested longitudinal commands ≤ -0.25, each at speed ≥ 20 km/h), the fault applies every negative longitudinal command at 70% of its requested value:

$$u_t^{\text{applied}} = \begin{cases} \beta u_t, & t \geq t_e,\ u_t < 0, \\ u_t, & \text{otherwise}, \end{cases} \qquad \beta = .70. \tag{19}$$

The controller receives no fault label; its policy is unchanged, and its steering and braking proposals respond to the altered state. Paired runs share configuration and seeds. This command attenuation emulates lost braking authority at the actuator interface; it does not model brake wear. Anchoring the onset to the first sustained braking pairs runs at a common index and emulates a braking-authority loss present when braking is first needed, not a fault beginning at an arbitrary time.

A scene qualifies ($E_i = 1$) only if the intact run arrives, completes at least 98% of the route, has no collision or road departure, and brakes firmly (an onset t_e exists, some request ≤ -0.50, and at least 5% of steps ≤ -0.25). *Intact* and *nominal* both denote the unfaulted condition. Physical failure Y_i is any collision or road departure; we estimate the fault's closed-loop effect on this selected population, $\Delta_E = \mathbb{E}[Y_i(.70) - Y_i(1.00) \mid E_i = 1]$. The effect repeats on disjoint cohorts: 11/60 and 10/60 fault failures, versus 0/60 intact failures in each (zero by selection); the pooled increase is 17.5 percentage points (pp; descriptive paired-bootstrap interval $[10.8, 24.2]$). Three onset-anchored measures, fixed before this exploratory comparison on exposed cohorts, do not clearly predict which fault drives fail (all area-under-ROC-curve, AUROC, intervals include .5; Appendix C.15).

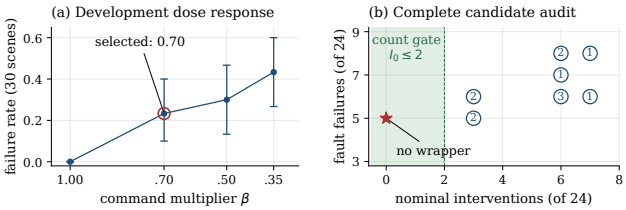

Figure 1: (a) Development dose response used to select β; bars are the paired-bootstrap intervals of Table 10 (intact failures are zero). (b) Frozen audit: I_0 counts intact scenes with a released fallback, shading marks the allowed $I_0 \leq 2$; numerals mark coincident candidates; "no wrapper" is the unmodified controller.

6.3 The frozen audit screen

Development selects $\beta = .70$ on 30 qualified scenes by a fixed rule: the largest tested multiplier with paired failure increase at least .10, a positive lower bootstrap endpoint, and strict success at least .50 (Figure 1a). The monitor predicts acceleration from the requested longitudinal command with a first-order lag model fit on nominal development braking and propagated recursively. When predicted acceleration is at most $-0.20\,\mathrm{m/s^2}$, it scores under-deceleration, positive observed-minus-predicted acceleration divided by $0.5\,\mathrm{m/s^2}$, either instantaneously or smoothed by an exponentially weighted moving average (EWMA). Each of the twelve frozen candidates, or *wrappers*, pairs one score, one threshold (the .95, .975, or .99 quantile of the 30 intact development scenes' maximum score in the second after t_e), and one latched fallback that, 0.5 s after the trigger, holds a constant emergency or service braking command (-6 or $-3\,\mathrm{m/s^2}$ requested) while keeping the policy's steering; the fallback does not affect triggering. Admission requires at most 2/24 nominal interventions and mean route completion at least 0.85 (Equation (18)). On 24 audit scenes, all candidates meet the completion requirement, but the least-triggering ones intervene in 3/24 intact scenes, so none is admitted (Figure 1b). This is a screen result, not a population-rate separation (Section 5.1). Counting only first crossings inside the one-second calibration window leaves the closest candidates at 3/24. Allowing a third intervention would pass four candidates with five or six fault failures, against five without a wrapper (Table 12).

A post-hoc contextual predictor adds speed and steering inputs with quadratic and interaction terms to the measured previous acceleration and requested command, fit by ridge regression; unlike the frozen monitor, it predicts one step ahead from measured acceleration rather than recursively. On the original predictor's 505 development braking intervals, in five seed-group folds with both models refit per fold, the original and contextual predictors have mean-fold root-mean-square error (RMSE) 2.230 and $1.390\,\mathrm{m/s^2}$ (37.7% lower; pooled out-of-fold, 36.5%), lower in all five folds. With the same gate, scaling, and EWMA, its EWMA thresholds detect 8–10/24 faults (original: 2–6/24) but still trigger in 3/24 intact audit scenes (original: 3–5/24), and its instantaneous thresholds trigger in 5–9/24. These are offline counts (one-second window, strict exceedance) on recorded no-wrapper traces, not closed-loop runs, and the ridge penalty was selected on the same folds (Appendix C.7).

6.4 Post-hoc response comparisons

After the frozen campaign, we compared responses on the same scenes (Table 2, Figure 2). Three braking responses share a trigger that applies the onset rule online, firing on the third of three consecutive qualifying braking requests. It fires in every drive: two decisions after t_e in intact drives and no earlier in fault drives, so it is aligned with the injected fault by construction rather than detecting it. *Emergency latch* then holds the emergency command to episode end; *emergency + release* holds it until speed stays at most 5 km/h for three decisions, then returns control; *full-brake latch* applies the maximum command -1 to episode end. *Command compensation* instead scales up the policy's braking commands. The oracle arm multiplies them by $1/.70$ from the known onset, in fault drives only; the fixed-gain arm applies the same gain from the common trigger in every drive. Twelve adaptive variants (C01–C12) apply gain $1/\widehat{\beta}$, clipped to $[1, 1/.70]$, from an online median $\widehat{\beta}$ of observed-to-predicted acceleration ratios, varying window, eligibility, and activation threshold (Appendix C.10); a distinct earlier adaptive policy is reported separately.

Table 2: Responses on the same 120 qualified scenes per condition. Rep./ind.: repaired/induced fault outcomes versus no wrapper; brackets: descriptive paired-bootstrap intervals. Nominal zeros: by selection (no wrapper), not applied (oracle).

Evaluation	Nominal fail	Fault fail	Rep./ind.	Paired change (pp)
Frozen no-wrapper stress contrast				
No wrapper	0/120	21/120	–	$+17.5$ $[10.8, 24.2]$
Post-hoc response contrasts: fault arm minus no-wrapper fault baseline				
Emergency latch	25/120	26/120	18/23	$+4.2$ $[-6.7, 14.2]$
Emergency + release	34/120	31/120	8/18	$+8.3$ $[0.0, 16.7]$
Full-brake latch	39/120	32/120	21/32	$+9.2$ $[-2.5, 20.8]$
Oracle compensation	0/120	8/120	15/2	-10.8
Fixed-gain compensation	14/120	16/120	10/5	-4.2
Earlier adaptive policy	8/120	24/120	4/7	$+2.5$
Adaptive, 12 variants	0–5/120	17–21/120	1–4/0–3	-3.3 to 0.0

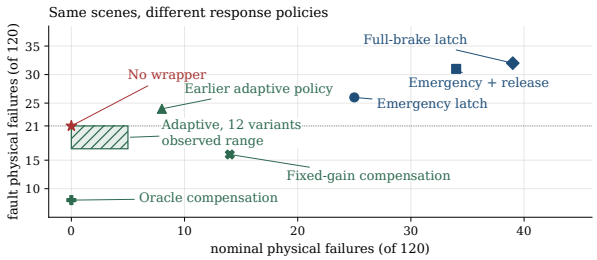

Figure 2: Fault versus nominal failures by response (120 scenes per condition). The hatched box spans C01–C12, not a confidence region; the oracle uses known fault timing and size.

Braking responses. Emergency latch repairs 18 baseline fault failures and induces 23; full-brake latch repairs 21 and induces 32. Emergency + release repairs eight and induces 18, while cutting nominal drives that never arrive from 113/120 under emergency latch to 10/120. With paired outcomes, the net failure-rate change is (induced − repaired)/120. Induced failures outnumber repairs in all three point estimates, but each paired interval includes or touches zero (Appendix C.9).

Command compensation. With known fault timing and size, the oracle arm yields 8/120 fault failures: 15 repaired and two induced. It is a privileged reference, not an achievable policy. The fixed-gain arm yields 16/120, with 14/120 nominal failures; with an onset-aligned trigger and the true gain, it is not a test of fault detection. An initial adaptive policy yields 24/120 fault failures; twelve later variants (C01–C12) yield 17–21/120 with 23–64 nominal interventions. None meets the post-hoc screen for carrying a variant forward to a fresh cohort, at most 12/120 nominal interventions and 16/120 fault failures, the latter the fixed-gain result (Appendices C.10, C.13). They leave 12–15/21 failures without activation before the first adverse event and activate in 23–64/120 intact runs; their ratio estimate has AUROC .645 $[.621, .676]$ under the permissive eligibility (Appendix C.14).

7 Related Work and Conclusion

Our ingredients come from partial observability and identification (Kaelbling et al., 1998; Manski, 2003; Pearl, 2013; Le Cam & Yang, 2000), change detection and fault-tolerant control (Page, 1954; Lorden, 1971; Tartakovsky et al., 2014; Chen & Patton, 1999; Isermann, 2006; Blanke et al., 2016), and runtime assurance and predictive safety (Seto et al., 1998; Ames et al., 2017; Bansal et al., 2017; Alshiekh et al., 2018; Wabersich & Zeilinger, 2023; Contreras et al., 2025); Appendix D compares them. Seeing danger and responding safely are different assurance achievements. The composed bound makes this separation explicit across decision ambiguity, consequence-model validity, timely detectability, and response execution; the MetaDrive study does not validate the bound, but illustrates why success on one obligation does not establish the others. Guarantees are premise-dependent; the evidence covers one controller, simulator, and fault family on selected scenes (Appendix C.17).

REPRODUCIBILITY STATEMENT

The formal results include explicit assumptions and proofs; Appendix A provides additional details. Appendix B specifies the synthetic models, and Appendix C records the controller, simulator configuration, qualification rules, monitor implementation, and full audit results. Appendix E identifies the archived programs, numerical records, and analysis utilities, and distinguishes the manuscript-source bundle from the separate trajectory archive.

ETHICS STATEMENT

The experiments use controlled simulation, with no human subjects or physical deployment. Their conditional failure rates are not real-road injury-risk estimates. No frozen wrapper was deployed on either final evaluation cohort after the operational audit rejected the candidate family. The later same-scene braking and command-compensation response studies are explicitly labeled post hoc; Appendix C.17 consolidates their evidence scope.

AI USE STATEMENT

Generative AI tools assisted with validating mathematical claims and proofs, experimental coding and analysis, and manuscript proofreading. The authors conceived the ideas, directed the experiments, checked every proof and every reported number against the archived result files, and retain full responsibility for all claims, proofs, code, numerical results, and submitted materials.

REFERENCES

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11797.

Aaron D. Ames, Xiangru Xu, Jessy W. Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8): 3861–3876, 2017. doi: 10.1109/TAC.2016.2638961.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J. Tomlin. Hamilton–Jacobi reachability: A brief overview and recent advances. In *IEEE Conference on Decision and Control*, pp. 2242–2253, 2017. doi: 10.1109/CDC.2017.8263977.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021.

Mogens Blanke, Michel Kinnaert, Jan Lunze, and Marcel Staroswiecki. *Diagnosis and Fault-Tolerant Control*. Springer, 3 edition, 2016. doi: 10.1007/978-3-662-47943-8.

Haoyuan Cai, Zhenghao Peng, and Bolei Zhou. Robot-gated interactive imitation learning with adaptive intervention mechanism. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6243–6256, 2025.

Jie Chen and Ron J. Patton. *Robust Model-Based Fault Diagnosis for Dynamic Systems*. Springer, 1999. doi: 10.1007/978-1-4615-5149-2.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Jose Leopoldo Contreras, Ola Shorinwa, and Mac Schwager. Safe, out-of-distribution-adaptive MPC with conformalized neural network ensembles. In *Proceedings of the 7th Annual Learning for Dynamics and Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pp. 194–207, 2025.

David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. arXiv preprint arXiv:2405.06624, 2024.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems*, 2022.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

International Organization for Standardization. ISO 26262:2018: Road vehicles—functional safety. ISO, 2018.

International Organization for Standardization. ISO 21448:2022: Road vehicles—safety of the intended functionality. ISO, 2022.

Rolf Isermann. *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer, 2006. doi: 10.1007/3-540-30368-5.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2 edition, 2000.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3461–3475, 2023. doi: 10.1109/TPAMI.2022.3190471. Published online in 2022.

Lars Lindemann, Xin Qin, Jyotirmoy V. Deshmukh, and George J. Pappas. Conformal prediction for STL runtime verification. In *ACM/IEEE International Conference on Cyber-Physical Systems*, pp. 142–153, 2023. doi: 10.1145/3576841.3585927.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971. doi: 10.1214/aoms/1177693055.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Judea Pearl. The mathematics of causal inference. Technical Report R-416, UCLA Cognitive Systems Laboratory, 2013.

John H. Reif. The complexity of two-player games of incomplete information. *Journal of Computer and System Sciences*, 29(2):274–301, 1984. doi: 10.1016/0022-0000(84)90034-5.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

Danbing Seto, Bruce H. Krogh, Lui Sha, and Alongkrit Chutinan. The Simplex architecture for safe on-line control system upgrades. In *Proceedings of the American Control Conference*, volume 6, pp. 3504–3508, 1998. doi: 10.1109/ACC.1998.703255.

Lui Sha. Using simplicity to control complexity. *IEEE Software*, 18(4):20–28, 2001. doi: 10.1109/MS.2001.936213.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. On a formal model of safe and scalable self-driving cars. arXiv:1708.06374, 2017.

Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman and Hall/CRC, 2014. doi: 10.1201/b17279.

UL Standards & Engagement. UL 4600: Standard for safety for the evaluation of autonomous products. UL Standards & Engagement, 2023.

Kim P. Wabersich and Melanie N. Zeilinger. Predictive control barrier functions: Enhanced safety mechanisms for learning-based control. *IEEE Transactions on Automatic Control*, 68(5):2638–2651, 2023. doi: 10.1109/TAC.2022.3175628.

Yiqi Zhao, Bardh Hoxha, Georgios Fainekos, Jyotirmoy V. Deshmukh, and Lars Lindemann. Robust conformal prediction for STL runtime verification under distribution shift. In *ACM/IEEE International Conference on Cyber-Physical Systems*, pp. 169–179, 2024. doi: 10.1109/ICCPS61052.2024.00022.

APPENDIX CONTENTS

A Further proofs, assumptions, and counterexamples **13**

A.1 Risk bounds on a latent covered stratum 13

A.2 Finite-audit actionwise validity . . . 13

A.3 Approximate non-identification . . . 14

A.4 No uniform internal certificate across indistinguishable environments 14

A.5 Risk identification is not low risk . . 14

A.6 Certification failure and physical failure have only a one-way relation . 14

A.7 Confidence-qualified finite-candidate admission 14

A.8 A worked stopping-envelope derivation 15

A.9 Deduplicated execution is not additional sample size 15

A.10 An exact worked example of the composed bound 15

A.11 Proofs of foundational propositions . 16

A.12 Standalone fallback accounting . . . 16

A.13 A finite-sample upper certificate from auditable cells 17

A.14 Conditional stagewise composition . 17

A.15 Sampling precision and action-consequence information 17

A.16 A nominal cost bound for response-sensitive admission 18

B Synthetic experiments and their targets **18**

B.1 Blind fork, triple conflict, and model error 18

B.2 A positive remedy: the occluded-crosswalk envelope 19

B.3 The declared 48-completion class . . 20

B.4 Composed bound for the archived 48-completion endpoint 20

B.5 Finite-audit prediction and its coverage statement 21

B.6 Mechanism holdout is not the same experiment 21

C MetaDrive specification and complete audit table **22**

C.1 Controller and execution configuration 22

C.2 Qualification and cohort accounting . 22

C.3 Development severity selection . . . 23

C.4 Monitor and fallback specification . . 23

C.5 Complete clean-audit result 24

C.6 Post-hoc calibration-window alignment diagnostic 24

C.7 Post-hoc contextual-monitor sensitivity test 25

C.8 Final no-wrapper campaign summary 26

C.9 Post-hoc braking-demand fallback frontier 26

C.10 Post-hoc compensation response study 27

C.11 Statistical interpretation and archived reference values 28

C.12 Empirical separation of the recorded alarm channel 29

C.13 Analysis lineage and carry-forward rule 29

C.14 Evidence budget of the frozen adaptive estimator 30

C.15 Matched failure-case trajectory diagnostic 31

C.16 Matched development-CV result . . 31

C.17 Limitations and evidence scope . . . 32

D Expanded related work **33**

E Experimental artifacts and reproducibility **34**

E.1 Programs and numerical records . . . 34

E.2 Experiment versioning and verification 35

A Further proofs, assumptions, and counterexamples

A.1 Risk bounds on a latent covered stratum

Let $C_t^\star$ indicate that the active consequence contract is actually valid, not merely declared valid by the monitor. If $C_t^\star$ is not measurable from H_t, unconditional action-risk approximation does not automatically hold after conditioning on coverage. The necessary premise is

$$\sup_a |L_P^{\mathrm{cov}}(a \mid h) - L_{\mathcal{M}_t}(a \mid h)| \le \delta_t, \quad L_P^{\mathrm{cov}}(a \mid h) = \mathbb{E}_P[\ell_t(a) \mid H_t = h, C_t^\star = 1]. \tag{20}$$

Let K_t indicate commitment and assume K_t and the chosen action a_t are history-measurable, including independent policy randomness. Define the optimization gap $\gamma_t = L_{\mathcal{M}_t}(a_t \mid H_t) - \rho_{\mathcal{M}_t}(H_t) \ge 0$. On the covered stratum,

$$(\rho_{\mathcal{M}_t} - \delta_t)_+ \le L_P^{\mathrm{cov}}(a_t \mid H_t) \le \rho_{\mathcal{M}_t} + \gamma_t + \delta_t. \tag{21}$$

The tower property gives

$$\mathbb{E}[\ell_t K_t C_t^\star] = \mathbb{E}[K_t \mathbb{P}(C_t^\star = 1 \mid H_t) L_P^{\mathrm{cov}}(a_t \mid H_t)] \tag{22}$$

$$\le \mathbb{E}[K_t C_t^\star (\rho_{\mathcal{M}_t} + \gamma_t + \delta_t)]. \tag{23}$$

Thus, with $M_\rho = \sum_t \mathbb{E}[K_t C_t^\star \rho_{\mathcal{M}_t}]$,

$$J_N = \mathbb{E} \sum_{t \le N} \ell_t K_t \le M_\rho + \sum_t \mathbb{E}[K_t C_t^\star (\gamma_t + \delta_t)] + J_N^{\mathrm{escape}}, \tag{24}$$

where

$$J_N^{\mathrm{escape}} = \sum_t \mathbb{E}[\ell_t K_t (1 - C_t^\star)] \le \sum_t \mathbb{P}(K_t = 1, C_t^\star = 0). \tag{25}$$

This separates covered ambiguity, optimization, model approximation, and uncovered loss without assuming independence between them. It is an accounting bound, not an estimator of the uncovered probability.

Why latent conditioning matters. Take constant H, $Y \sim \mathrm{Bernoulli}(1/2)$, and $L_{\mathcal{M}} = \mathbb{E}Y = 1/2$. Unconditional prediction is exact. For the latent event $C = Y$, however, $\mathbb{E}[YC] = 1/2$ while $\mathbb{E}[L_{\mathcal{M}} C] = 1/4$. This counterexample explains why an unconditional validity claim cannot be silently multiplied by an arbitrary latent coverage event.

A.2 Finite-audit actionwise validity

For a finite action family, obtain n independent bounded losses from the same declared operational stratum for every action. Let $\widehat{L}_P(a)$ be each empirical mean and $L_{\mathcal{M}}(a)$ the prediction averaged over that same stratum. Hoeffding and a union bound give, with probability at least $1 - \kappa$,

$$\sup_a |L_P(a) - L_{\mathcal{M}}(a)| \le \sup_a |\widehat{L}_P(a) - L_{\mathcal{M}}(a)| + \sqrt{\frac{\log(2|\mathcal{A}|/\kappa)}{2n}}. \tag{26}$$

The triangle inequality supplies the final step. This is a stratum-average certificate. It is not a history-uniform guarantee or a guarantee conditional on a hidden coverage label unless additional assumptions establish those stronger statements. Merely increasing rollout count does not audit actions that were never sampled.

A.3 APPROXIMATE NON-IDENTIFICATION

Suppose two environments P_0, P_1 have observable laws $Q_i = Q_{P_i}^{\mathcal{I}}$ with $\mathrm{TV}(Q_0, Q_1) \le \varepsilon$, while their fixed-policy risks differ by Δ_R. For any report $g(H) \in [0,1]$, write $g_i = \mathbb{E}_{Q_i} g(H)$ and $r_i = R_{P_i}(\pi)$. Then

$$\Delta_R = |r_1 - r_0| \le |r_1 - g_1| + |g_1 - g_0| + |g_0 - r_0| \le 2 \max_i |r_i - g_i| + \varepsilon. \tag{27}$$

Therefore the worst marginal bias is at least $(\Delta_R - \varepsilon)_+/2$. This result concerns a fixed information-collection protocol. If a proposed active probe changes the observable laws, its safety, time cost, and new laws must be analyzed explicitly.

A.4 NO UNIFORM INTERNAL CERTIFICATE ACROSS INDISTINGUISHABLE ENVIRONMENTS

Let two environments induce the same law Q of available history and active model but have true validity labels $V_0(h), V_1(h)$ that disagree on a set of positive Q probability. For every common declaration $\widehat{V}(h)$,

$$\mathbb{P}_0(\widehat{V} \ne V_0) + \mathbb{P}_1(\widehat{V} \ne V_1) \ge Q(V_0 \ne V_1) > 0. \tag{28}$$

At each disagreeing history a single binary answer cannot equal both labels; integration proves the inequality. The same argument applies after incorporating independent randomization. This does not say that validity for a fixed known P is nonmeasurable. It says the same observation-only implementation cannot be almost surely correct in both indistinguishable environments without further evidence or restrictions.

A.5 RISK IDENTIFICATION IS NOT LOW RISK

A singleton identified set $\{0.9\}$ has diameter zero but certifies high risk. Conversely, a blanket safe action can turn policy-risk set $[0,1]$ into $\{0\}$ while leaving the observation law unchanged. A causal diagnosis is therefore neither necessary for every protective action nor sufficient for every safe action: one can protect without identifying, or identify that every remaining action is dangerous.

A.6 CERTIFICATION FAILURE AND PHYSICAL FAILURE HAVE ONLY A ONE-WAY RELATION

Suppose a fallback contract ensures no loss whenever restoration occurs by t_{cert}. The implication is

$$\{\text{valid contract and } \tau_r \le t_{\mathrm{cert}}\} \subseteq F^c. \tag{29}$$

It does not imply $\{\tau_r > t_{\mathrm{cert}}\} \subseteq F$. A stronger actual actuator may still stop after the conservative cutoff. Nor does a first alarm at t_{cert} suffice when actuation begins later. The discriminator's information window must leave the required physical latency, and the certificate must be valid for the true state during entry and execution.

A.7 CONFIDENCE-QUALIFIED FINITE-CANDIDATE ADMISSION

The empirical MetaDrive gate in Equation (18) was deliberately not interpreted as a population confidence guarantee. For completeness, a stronger analytical refinement is:

Proposition A.1 (Confidence-qualified finite-candidate admission)**.** *Condition on a development-frozen set of M candidates. For each of n independent audit scenes, observe bounded nominal intervention $I_i(\theta) \in [0,1]$ and bounded progress $U_i(\theta) \in [0,1]$. Their joint vector may be dependent across candidates and metrics, but scenes must be independent and representative of the claimed population. With*

$$h_n = \sqrt{\frac{\log(2M/\kappa)}{2n}}, \quad \widehat{\Theta}_{\mathrm{conf}} = \{\theta : \widehat{I}(\theta) + h_n \le \bar{I},\ \widehat{U}(\theta) - h_n \ge \bar{U}\}, \tag{30}$$

with probability at least $1 - \kappa$, every admitted candidate satisfies $I(\theta) \le \bar{I}$ and $U(\theta) \ge \bar{U}$. Any audit-dependent selector restricted to this set preserves that statement.

Proof. For each candidate, one-sided Hoeffding bounds control $I - \widehat{I}$ and $\widehat{U} - U$. A union bound over $2M$ events gives probability at least $1 - \kappa$ that all deviations are at most h_n. Every member of the displayed set then satisfies both population inequalities simultaneously, so selecting among them cannot invalidate the event. □

The bound certifies only its two operational constraints; fallback consequence validity and physical-risk control remain separate obligations. For $M = 12$, $n = 24$, and $\kappa = 0.05$, $h_n \approx 0.359$, so this conservative uniform bound cannot certify a 10% intervention limit even with no observed interventions. More efficient exact or variance-sensitive bounds may be used under their assumptions, but their certification claim must not be retroactively attributed to the executed empirical gate.

For an exact single-indicator rule, audit size is part of the design. For one fixed Bernoulli indicator, the exact one-sided 95% upper limit after $0/24$ is $1 - 0.05^{1/24} \approx 0.1173$. A 10% limit can first be certified with at most $0, 1, 2, 3, 4$ interventions at $n = 29, 46, 61, 76, 89$, respectively (Clopper & Pearson, 1934). These single-candidate limits distinguish screening from certification. The simultaneous finite-candidate refinement above addresses selection; scenes, rather than timesteps or reused trajectories, are the sampling units.

A.8 A WORKED STOPPING-ENVELOPE DERIVATION

Before braking, constant speed over latency L consumes distance vL. Thereafter $\dot{v} = -b$ with $b \geq \underline{b}$ and no reverse motion implies stopping distance at most $v^2/(2\underline{b})$. Adding the two yields Equation (1). Multiplying by $2\underline{b}$ and solving the quadratic inequality for nonnegative v yields Equation (15). The proof assumes known separation, a stationary hazard in the relevant direction, bounded delay, and the stated deceleration floor. It does not address rear impact, steering constraints, road curvature, or uncertain traction unless those are incorporated in the contract.

A.9 DEDUPLICATED EXECUTION IS NOT ADDITIONAL SAMPLE SIZE

The candidate implementation triggers once, queues a single command, and thereafter permanently overrides the longitudinal action. Under a deterministic simulator/controller implementation with common seeds and no monitor side effects, the physical trajectory equals the no-wrapper trajectory until actual fallback release. If no command is released before episode termination, its physical outcome is the baseline outcome. If two logical candidates share scene, condition, release step, and fallback action, they share all applied actions and subsequent states by induction. They may have different recorded alarm times, which must remain distinct in logical records.

This justifies computational reuse only under the stated implementation equivalence and numerical reproducibility. It does not turn 70 physical executions into 576 independent observations: each candidate is evaluated on 24 distinct paired scenes, and duplicated trajectory references create no new statistical sample. The new campaign reported the logical-to-physical execution mapping; independent verification of all equivalence prefixes would require the raw traces and exact executed source versions.

A.10 AN EXACT WORKED EXAMPLE OF THE COMPOSED BOUND

This finite construction illustrates Theorem 5.2; it is not a new simulator result. Take histories h_1, h_2, h_3, $Q_C = (0.6, 0.3, 0.1)$, $Q_E = (0.1, 0.3, 0.6)$, escape mass $e = 0.2$, and response event $D = \{h_3\}$. Let $\rho_{\mathcal{M}} = (0.04, 0.06, 0.08)$, $\gamma = 0.01$, and $\delta = 0.02$. A two-action primary family has modeled losses $\rho_{\mathcal{M}}$ and $\rho_{\mathcal{M}} + 0.01$ and the same covered-stratum true losses; the policy chooses the second. On the uncovered stratum its primary loss probability is 0.7. Conditional on invoking a response, G occurs with probability 0.9; response loss probability is 0.02 on G and 1 on G^c. These specifications define a finite Bernoulli-loss experiment with $\epsilon_f = 0.02$ and a valid missed-response lower bound $\phi = 0.5$.

The values are $\alpha = 0.1$, $m = 0.4$, $\mathrm{TV}(Q_C, Q_E) = 0.5$, $\mathbb{P}(D) = 0.2$, $\xi = 0.02$, $A_- = 0.0192$, and $A_+ = 0.0552$. The risk decomposition evaluates to

$$R_\pi = \underbrace{0.0408}_{C \cap D^c} + \underbrace{0.0560}_{C^c \cap D^c} + \underbrace{0.0236}_{D} = 0.1204, \qquad 0.0592 \leq R_\pi \leq 0.1588.$$

The lower endpoint uses the declared ϕ, not an inference from alarm timing. The upper endpoint uses the actual miss probability m, not a lower bound on history separation. The accompanying arithmetic check enumerates all eight response events for these fixed laws and verifies the composed inequalities exactly with rational arithmetic. The proof above establishes the general statement; the finite calculation is a reproducible illustration.

A.11 Proofs of foundational propositions

Proof of Proposition 2.3 (Covered risk and approximate-model transfer).

Proof. Uniform validity extends to action mixtures by averaging. Their model risk is at least ρ, so their true risk is at least $\rho - \delta$, and nonnegativity gives the positive part. At $\hat{a}$, model risk is exactly ρ, proving the first transfer statement. Next, $L_P(\hat{a}) \le L_{\mathcal{M}}(\hat{a}) + \delta \le L_{\mathcal{M}}(a^\star) + \delta \le L_P(a^\star) + 2\delta$. For the final equivalence, a common safe action has zero expectation. Conversely, a zero-risk minimizing action has nonnegative loss summands whose positive-belief weighted sum is zero, hence all those summands vanish. A minimizing mixture must put its positive mass on such zero-risk actions. □

Proof of Proposition 2.4 (Value of timely, valid information).

Proof. After receiving X, one can ignore it and take an action optimal under H. Optimizing separately for each refined posterior cannot have greater conditional expected loss than that feasible policy. □

Proof of Proposition 3.1 (Contract-relative risk-reporting limit).

Proof. Write $r_- = \inf \Gamma_\pi(c)$ and $r_+ = \sup \Gamma_\pi(c)$. For any two attainable risks r_1, r_2, the triangle inequality gives $2\max(|g - r_1|, |g - r_2|) \ge |r_1 - r_2|$. Approaching the two endpoints proves the lower bound, and their midpoint has distance at most $(r_+ - r_-)/2$ from every attainable risk. An upper certificate must dominate every attainable risk and hence their supremum. Finally, $\mathbb{E}_P g(H)$ is identical under the shared observable law, so the scalar argument applies to that common expectation. □

Proof of Proposition 4.1 (Deadline-limited monitorability frontier).

Proof. By the variational definition of total variation, $Q_1(D) - Q_0(D) \le \mathrm{TV}(Q_0, Q_1)$. Subtracting from one gives Equation (11). Selecting $D = \{q_1 > q_0\}$ attains equality. For unequal costs, at each history select the smaller of the two contributions; $D = \{w_1 q_1 > w_0 q_0\}$ gives the stated integral. Independent randomization can be incorporated in the history and cannot improve this pointwise minimum. □

A.12 Standalone fallback accounting

Let G denote the event that the selected fallback's consequence contract holds and effective restoration occurs by its certified cutoff. Suppose the conditional failure risk on G is at most ϵ_f.

Proposition A.2 (Fallback assurance budget)**.** *If $\mathbb{P}(G^c) \le \eta$, the specified failure probability satisfies*

$$\mathbb{P}(F) \le \epsilon_f + (1 - \epsilon_f)\eta. \tag{31}$$

In particular, a valid, reachable, zero-loss fallback that becomes effective in time prevents the specified loss. If invalidity, failure to alarm in time, and excess actuation delay have probabilities at most η_v, η_d, η_l, and these cover G^c, then one may take $\eta \le \min(1, \eta_v + \eta_d + \eta_l)$ without assuming independence. In Theorem 5.2, a late alarm belongs to D^c, so only η_v and η_l remain.

Proof. Split failure on G and G^c: $\mathbb{P}(F) \le \epsilon_f \mathbb{P}(G) + \mathbb{P}(G^c) = \epsilon_f + (1 - \epsilon_f)\mathbb{P}(G^c)$. The final bound follows from the union bound on the stated failure events. □

This budget organizes the evidence needed for a response: a consequence-risk bound on the valid execution event and separate bounds on the ways that event can fail. The admission record therefore includes response validity and execution evidence alongside monitor performance. Where the evidence is insufficient, the record identifies the outstanding obligation and withholds promotion of that candidate.

A.13 A FINITE-SAMPLE UPPER CERTIFICATE FROM AUDITABLE CELLS

The upper endpoint of Theorem 5.2 becomes an estimable object when the covered stratum C, response-success event G, and a covered chosen-action envelope $u(H) = \min\{1, \rho_{\mathcal{M}}(H) + \gamma(H) + \delta(H)\}$ have independent justification. Define

$$W = \mathbf{1}_C \mathbf{1}_{D^c} u(H) + \mathbf{1}_{C^c} \mathbf{1}_{D^c} + \epsilon_f \mathbf{1}_D \mathbf{1}_G + \mathbf{1}_D \mathbf{1}_{G^c}. \tag{32}$$

The four disjoint cells give $0 \le W \le 1$ and $R_\pi \le \mathbb{E}W$. For a development-fixed family of M candidates and n independent representative labeled audit units, Hoeffding and a union bound imply, with probability at least $1 - \kappa$,

$$\forall \theta : \quad R_{\pi_\theta} \le \min\left\{1, \overline{W}_\theta + \sqrt{\frac{\log(M/\kappa)}{2n}}\right\}. \tag{33}$$

Candidate scores on the same unit may be dependent. The development-fixed contract and ϵ_f must be valid before this sampling statement; if estimated from separate data, their failure probabilities are added to κ. Audit-dependent selection within the simultaneous bounds preserves coverage. This construction estimates escaped misses jointly, through $\mathbf{1}_{C^c}\mathbf{1}_{D^c}$, rather than inferring an upper miss rate from a TV lower witness. A latent coverage label without independent audit evidence cannot be substituted into the score. This is a prospective certificate, not an executed MetaDrive guarantee.

A.14 CONDITIONAL STAGEWISE COMPOSITION

Let $\mathcal{H}_t^-$ contain the actual history before stage t, including preceding interventions. At that stage define a bounded incremental loss $Y_{\pi,t}$ and a common local pre-response protocol, then impose Theorem 5.2's premises conditional on $\mathcal{H}_t^-$. Let $U_t(\mathcal{H}_t^-)$ be its conditional upper endpoint. The tower property gives

$$\mathbb{E}\left[\sum_{t=1}^{N} Y_{\pi,t}\right] \le \sum_{t=1}^{N} \mathbb{E}[U_t(\mathcal{H}_t^-)]. \tag{34}$$

No independence between stages is needed. For episode failure, use disjoint first-failure increments $Y_{\pi,t} = \mathbf{1}\{T_F = t\}$, so the sum is the any-failure indicator. For ordinary repeated adverse flags the sum instead bounds an event count, and a union bound gives an upper bound on any failure. Stage contracts refer to local loss and their specified response segments; repeated use of overlapping full-episode losses is not this construction. A latched policy continues in the state history after activation, with no assumption that later states follow the no-wrapper law. This is the sequential extension of Appendix A.1's time-indexed accounting, with additional per-stage response premises; the archived MetaDrive endpoints alone do not calibrate these premises.

A.15 SAMPLING PRECISION AND ACTION-CONSEQUENCE INFORMATION

Section 3 separates uncertainty within a declared class from uncertainty about that class. The finite audit quantifies the former:

Proposition A.3 (Finite audit of a declared class)**.** *For a fixed policy, suppose an external audit obtains N independent losses in $[0,1]$ from each of K declared completions, with a correct declared law-cell partition. With probability at least $1 - \kappa$,*

$$\max_k |\widehat{r}_k - r_k| \le \epsilon_N, \quad \max_c |\widehat{\mathrm{diam}}_\pi(c) - \mathrm{diam}_\pi(c)| \le 2\epsilon_N, \quad \epsilon_N = \sqrt{\frac{\log(2K/\kappa)}{2N}}. \tag{35}$$

Proof. Hoeffding's inequality (Hoeffding, 1963) and a union bound control all K risks by ϵ_N. On that event the maximum and minimum over any fixed cell each change by at most ϵ_N, so their difference changes by at most $2\epsilon_N$. □

A safe blanket response can reduce policy risk without improving information about the consequences of the full action family. For a fixed family and a fixed observation protocol, that information is described by

$$\operatorname{diam}_{\mathcal{I}}^{A}(c) = \sup_{P,Q \in c} \sup_{a \in \mathcal{A}} |\mathbb{E}_P \ell(\omega, a) - \mathbb{E}_Q \ell(\omega, a)| \tag{36}$$

The policy-specific diameter diam_π and this action-family diameter answer different questions. The first can shrink because the selected policy is safer; the second holds the action family fixed while assessing the available information.

A.16 A NOMINAL COST BOUND FOR RESPONSE-SENSITIVE ADMISSION

A count budget assigns every intervention episode the same cost. A prospective alternative declares a nominal response cost $c_0(h, a)$ and constrains $\mathbb{E}_0[c_0(H, a(H))]$, retaining a separate consequence-risk requirement. A useful choice upper-bounds the positive increase in nominal loss. For $a \neq a_0$, suppose response and baseline share an actionwise δ-valid nominal contract. Then

$$[L_P(a \mid h) - L_P(a_0 \mid h)]_+ \leq c_0(h, a) := \min\{1, [L_{\mathcal{M}}(a \mid h) - L_{\mathcal{M}}(a_0 \mid h) + 2\delta]_+\}. \tag{37}$$

Indeed, the two validity inequalities give $L_P(a \mid h) - L_P(a_0 \mid h) \leq L_{\mathcal{M}}(a \mid h) - L_{\mathcal{M}}(a_0 \mid h) + 2\delta$. Taking the positive part and using the unit loss bound proves the displayed inequality. Set $c_0(h, a_0) = 0$ for no intervention. Other operational costs require declared units and weights. The cost definition and its supporting validity evidence must precede the audit; this construction does not retrospectively replace the campaign's count gate.

B SYNTHETIC EXPERIMENTS AND THEIR TARGETS

Table 3: Claim–evidence correspondence. The studies make different parts of the assurance argument measurable.

Obligation	Experiment	Supported result
Covered risk; preserve options	Blind fork, triple conflict, crosswalk dynamic program	Analytic floors and safe progress within the declared model
Risk sets; finite evidence	Declared 48-completion class; audit/test holdout design	Exact risk sets and composed bounds; audit design only (Appendix B.5)
Timely monitoring	Exact history laws and mechanism holdout	Channel- and cutoff-relative error frontier
Fallback consequences	Specified response model and exact stopping diagnostic	Protection conditional on validity and timing
Operational admission	Twelve candidates, 24 audit scenes	Complete implementation-level screening decision
Consequence sensitivity	Two disjoint 60-scene no-wrapper cohorts	Replicated conditional simulator stress effect
Response choice	Post-hoc braking and compensation studies	Paired repair, induction, and burden profiles

The synthetic programs and result JSON files are evaluated without resampling. The preservation manifest records the six scientific files used for the reported synthetic results. Numerical statements in this appendix use the experiment's original endpoint definitions.

B.1 BLIND FORK, TRIPLE CONFLICT, AND MODEL ERROR

In the blind-fork model an obstacle lies $D = 60\,\mathrm{m}$ ahead, the resolving observation time is exponential with rate $\lambda = 0.35\,\mathrm{s}^{-1}$, and a common safe shoulder independently exists with probability $p_{\mathrm{sh}} = 0.25$. At approach speed v, the mass of unresolved commitments without a safe shoulder is

$$\mu(v) = e^{-\lambda D/v}(1 - p_{\mathrm{sh}}). \tag{38}$$

For hidden-lane prior $(b, 1-b)$ and wrong-choice severity ϕ_0, minimum expected loss is $\mu(v)\min(b, 1-b)\phi_0$. The implemented optimal policy uses a resolving observation when available, otherwise the common shoulder, otherwise the lower-risk lane. Its maximum deviation from the analytic floor is 0.00197; a suboptimal ("impatient") policy's minimum excess is 0.01792. An oracle observing hidden state violates the information restriction and serves only as a sanity baseline.

For safe sets $\{a, b\}$, $\{b, c\}$, and $\{a, c\}$ with belief $(p, (1-p)/2, (1-p)/2)$, the action risks are $(1-p)/2, (1-p)/2, p$. Thus $\rho(p) = \min\{p, (1-p)/2\}$, with exact maximum $1/3$. The Monte Carlo grid has maximum error 0.00254; a grid that omits $p = 1/3$ need not attain the analytic maximum.

The model-error generator samples small finite decision problems, corrupts belief logits and binary consequence labels, and compares chosen-action predicted risk, true risk, and true optimum. Its heaviest tested corruption yields mean absolute risk error 0.145 and decision regret 0.089. These are empirical sensitivity measures, not independent certificates that a chosen uniform δ holds.

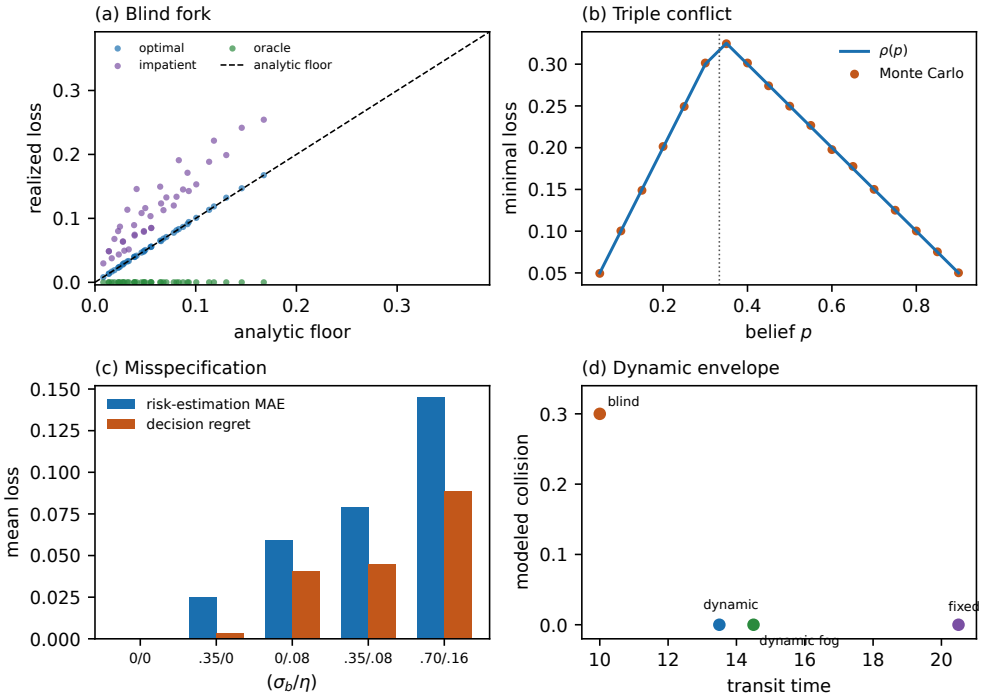

Figure 3: Covered-model decision-risk checks: analytic decision floors, triple conflict, misspecification sensitivity, and dynamic option preservation. In panel (c), σ_b and η are the belief-logit and consequence-label corruption levels (this η is unrelated to Remark 5.1). The claims are conditional on the declared models.

B.2 A POSITIVE REMEDY: THE OCCLUDED-CROSSWALK ENVELOPE

The discrete state is position, velocity, and knowledge of whether a crosswalk is clear or occupied. Actions change velocity subject to acceleration and braking limits. Before visibility resolves the crosswalk, admissible controls keep discrete stopping distance within the remaining protected region. Exact backward induction minimizes time over those controls. A fixed conservative speed cap belongs to the feasible policy class, so optimizing over the dynamic envelope cannot be slower under the same model and objective.

In the reported forty-cell corridor, pedestrian probability is 0.3, sight distance is six cells in clear weather and three in fog, maximum velocity is five, acceleration bound is one, and braking bound is two. The blind policy takes ten steps with collision probability 0.30. The fixed conservative speed cap takes 20.5 steps. The dynamic policy takes 13.5 in clear weather and 14.5 in fog while

respecting the zero-modeled-collision constraint. Within this declared model, preserving a stopping option is faster than the fixed cap with the same zero modeled collision probability.

B.3 THE DECLARED 48-COMPLETION CLASS

The class consists of six nominal completions, six benign visual novelties, twelve early sensor corruptions, twelve late brake degradations, and twelve silent brake failures. Family masses are $(0.65, 0.15, 0.10, 0.05, 0.05)$, uniform within each family. Nominal independent Bernoulli visual, disagreement, and health residual probabilities are $(0.010, 0.005, 0.005)$. Benign novelty changes the visual channel; early corruption changes visual/disagreement signals; late degradation changes health only after the common cutoff; the silent family matches all available pre-cutoff nominal laws. The tables below compare never- and always-trigger rules, a residual rule on a single residual channel (eight possible histories), an ensemble rule on the combined channels (512 histories), a health-channel rule, and an oracle reference that knows the hazard status.

The external audit receives completion-labeled samples and the declared partition by full signal law. The runtime monitor does not receive the completion label, true brake authority, or cell index. The test is sample holdout within a declared class, not open-ended discovery of previously unconceived physical mechanisms.

For the exact braking diagnostic, $x_h = 25$, initial velocity $v_0 = 10$, and $\Delta t = 0.25$. After trigger at step t,

$$v_{j+1} = [v_j - b\Delta t]_+, \qquad x_{j+1} = x_j + v_{j+1}\Delta t, \qquad x_t = v_0 t \Delta t. \tag{39}$$

The last-safe trigger steps for $b = 2.5, 3, 4, 5$ are $2, 3, 5, 6$. The common robust trigger cutoff is therefore step 2. For $b = 5$, triggering at step 3 stops at $16.25 < 25$: certification is missed, yet stopping succeeds.

The original archived loss is $Z_{\text{haz}}\mathbf{1}\{\tau_{\text{trig}} > 2\}$. It is not a collision label. The separate stopping diagnostic replaces step 2 with the completion-specific physical last-safe trigger. The original `false_confidence` field equals archived loss by construction and is not an independent test of true versus declared coverage.

Table 4: Exact monitorability in the declared 48-completion model. Both errors refer to triggering by the same declared cutoff.

Policy/channel	TV	$1 - \text{TV}$	Benign trigger	Hazardous miss	Error sum
Never trigger	.000	1.000	.000	1.000	1.000
Always trigger	.000	1.000	1.000	.000	1.000
Residual	.129	.871	.107	.764	.871
Ensemble	.313	.687	.046	.653	.700
Health	.000	1.000	.000	1.000	1.000
Oracle	1.000	.000	.000	.000	.000

Table 5: Policy-risk ambiguity is not information quality. Weighted diameters, certification risk, and assigned progress utility are defined by the experiment. Values are rounded.

Policy	Information diameter	Policy-risk diameter	Certification risk	Utility
Never trigger	1.000	1.000	.200	10.000
Always trigger	1.000	.000	.000	5.400
Residual	1.000	.867	.153	8.629
Ensemble	.750	.728	.131	8.520
Health	1.000	1.000	.200	9.846
Oracle	.000	.000	.000	9.080

B.4 COMPOSED BOUND FOR THE ARCHIVED 48-COMPLETION ENDPOINT

The archived loss is $Z_{\text{haz}}\mathbf{1}\{\tau_{\text{trig}} > 2\}$, so this specialization uses *certification loss*, not collision. Let $C = \{Z_{\text{haz}} = 0\}$, $D = \{\tau_{\text{trig}} \le 2\}$, $Y_0 = Z_{\text{haz}}$, and $Y_f = 0$. The benign family mass

is $.65 + .15 = .80$, hence $e = .20$. On C, both continuation and timely response have zero loss. Taking their exact zero-loss covered model gives $\rho_{\mathcal{M}} = \gamma = \delta = 0$ and $A_- = A_+ = 0$. Timely response is defined to avoid this endpoint, so $\epsilon_f = \xi = 0$ and $\phi = 1$. Theorem 5.2 becomes

$$.20[1 - \alpha - \mathrm{TV}(Q_C, Q_E)]_+ \leq R_\pi = .20m. \tag{40}$$

This is an exact specialization of the original endpoint. The monitorability table above reports α (benign trigger), m (hazardous miss), and TV rounded to three decimals. Treating each entry as an interval of half-width $.0005$ yields the following outward bounds; it does not recreate the unavailable unrounded 48-completion generator.

Table 6: Composed certification-loss bounds from the archived rounded law summaries. The interval in the last column uses $R_\pi = .20m$ directly; its width reflects numerical rounding, not sampling uncertainty.

Channel	Composed lower	Composed upper	R_π from miss rate
Residual	.1526	.1529	[.1527,.1529]
Ensemble	.1280	.1307	[.1305,.1307]

The residual rule attains the testing frontier to the reported precision; the ensemble rule has a nonzero gap. Using the displayed α, m, TV values gives lower $.1282$ and risk $.1306$, with the separately rounded error-sum column consistent with a gap of this order. Always-trigger and oracle have certification risk zero, while never-trigger and health have risk $.20$. Thus the monitoring term of the composed bound is instantiated on an experiment already in the paper; the covered and response terms are zero by the definition of this endpoint. It does not separately test physical fallback validity or latency. The independent finite example in Appendix A.10 exercises nonzero covered and failed-response terms. Script `verify_composed_bounds.py` checks this specialization and the rounding intervals using rational arithmetic.

B.5 Finite-audit prediction and its coverage statement

For each completion j in a declared law cell c, let $\widehat{r}_j$ be an audit estimate. The cell report is the midpoint of the audit extremes, and the prediction of its maximum error is $\widehat{D}_\pi(c)/2$. An independent test vector $\widetilde{r}$ scores $\max_{j \in c} |\widetilde{r}_j - \widehat{g}_c|$. Its difference from the audit-vector error is at most $\|\widetilde{r} - \widehat{r}\|_\infty$. Weighted Hoeffding applied to independent audit and test means gives a simultaneous bound

$$\|\widetilde{r} - \widehat{r}\|_\infty \leq \sqrt{\frac{\log(2K/\alpha)}{2}\left(\frac{1}{N_{\mathrm{audit}}} + \frac{1}{N_{\mathrm{test}}}\right)} \leq \epsilon_{\mathrm{audit}} + \epsilon_{\mathrm{test}}. \tag{41}$$

This justifies the audit/test band at one fixed policy and sample size. It is not a simultaneous guarantee over every policy, repetition, and sample-size choice unless those extra comparisons are included.

The study uses audit sizes $100, 250, 500, 1000, 2000, 5000$, 200 repetitions at each size, and $20{,}000$ test rollouts per completion. For monitor fitting, audit histograms estimate benign/hazardous history laws, the empirical likelihood-ratio rule is frozen, and evaluation uses $50{,}000$ new samples per class. The ensemble alphabet has 512 histories; the single residual channel has eight, so monitor estimation error can differ between them at equal audit sample size. This appendix specifies the design and its coverage statement; the paper does not report or rely on this study's numerical results.

B.6 Mechanism holdout is not the same experiment

The Gaussian mechanism-holdout study freezes a maximum-residual threshold from $4{,}000$ nominal nine-step episodes at a target 8% episode trigger rate. Mean shift motivates the statistic. Variance shift, late shift, and a silent (observation-matched) hazard are withheld until evaluation. Reported means and split standard deviations use 20 independent calibration/test splits and $20{,}000$ episodes per mechanism. The controlled loss assigns protection to timely intervention and therefore tests a stipulated consequence model; it does not empirically measure a real braking response.

Table 7: Mechanism-holdout results. The miss rate is the experiment-defined timing/consequence loss.

Mechanism	Trigger rate	Miss rate (loss)
Nominal	$.078 \pm .004$	0
Benign mean shift	$.979 \pm .002$	0
Hazardous design-visible mean shift	$.979 \pm .002$	$.021 \pm .002$
Withheld variance shift	$.610 \pm .008$	$.390 \pm .008$
Withheld late shift	$.512 \pm .010$	$.488 \pm .010$
Withheld silent hazard	$.079 \pm .004$	$.921 \pm .004$

C METADRIVE SPECIFICATION AND COMPLETE AUDIT TABLE

Stage names used in this appendix. Stage 4 is nominal development calibration; Stage 5A qualifies the 24 audit scenes; Stage 5B records paired no-wrapper audit runs; Stage 5C executes the frozen monitor/response candidates. Stages 6–7 select and evaluate the held-out cohort, and Stages 9–10 do the same for the disjoint replication cohort. These names identify archived execution stages, not additional independent datasets.

C.1 CONTROLLER AND EXECUTION CONFIGURATION

The evaluated checkpoint is the fixed PPO-trained actor released with AIM (Cai et al., 2025), as recorded in the campaign metadata and reproducibility package. Checkpoint inspection found 13 state-dictionary tensors and exact tensor equality between the original checkpoint and its compatibility conversion. The campaign executes the actor directly: two 64-unit tanh hidden layers, 259 input values, two action outputs, and clipping to $[-1, 1]$. The value network and policy variance do not determine deterministic mean actions. The stored archive name contains "20m"; its metadata reports `num_timesteps=9600000`; the recorded actor metadata, rather than the archive name, identifies the evaluated policy.

The execution stack is Python 3.13.15, NumPy 2.1.3, PyTorch 2.11.0+cu128, MetaDrive 0.4.3, Panda3D 1.10.15, and Panda3D-GLTF 0.13. The actor runs on CPU with one intra-process thread; four separate processes own separate simulator engines. This is the configuration actually used, not a claim that all versions satisfy MetaDrive's declared package metadata. Environment recreation must preserve the exact working configuration and assets rather than substitute newer dependencies silently.

Table 8: Frozen simulator configuration.

Setting	Value
Traffic density	0.06
Random traffic	false
Accident probability setting	0.8
Episode horizon	1,500 decisions
Physics step	0.02 s
Decision repeat	5
Decision interval	0.10 s
Vehicle/object crash terminates	false / false
Road departure / route departure terminates	true / true
Rendering / manual control	false / false
Run-seed rule	60,000,000 + scene seed
Reset seed	scene seed

C.2 QUALIFICATION AND COHORT ACCOUNTING

The nominal braking event starts at the first index $i \geq 20$ for which three consecutive requested longitudinal commands are at most -0.25 and each corresponding pre-action speed is at least $20\,\mathrm{km/h}$. Brake relevance additionally requires a minimum requested longitudinal command at most -0.50

and at least 5% of episode steps at or below -0.25. Strict success requires arrival, completion at least 0.98, and no recorded vehicle collision, object collision, or road departure. Cohort membership uses the intersection of strict success and brake relevance. Scenes in the 5000-series are reserved for a future fresh confirmatory cohort; no analysis in this paper opens, simulates, or scores them.

The event is extracted offline from the intact trajectory. It is available to the intervention generator and evaluator, not as an online oracle for the monitor. Both conditions use the frozen event index, rather than allowing the fault trajectory to choose its own onset.

Table 9: Scene accounting. Replays are integrity checks, not additional independent experimental units. Final parallel batches can include eligible scenes beyond the selected cutoff.

Role	Actual nominal range	Screened	Eligible found	Selected
Development	6000–6199	200	30	30
Audit	7000–7186	187	24	24
Held-out	8000–8399	400	64	60
Replication	9000–9319	320	70	60

Development qualification was replayed twice per selected scene and audit qualification once. Both final cohorts were replayed before paired no-wrapper evaluation. The selected cutoffs are 8370 and 9285. The held-out qualification replay matched all 60 stored trajectories exactly. The replication qualification replay checked discrete and numeric episode fields but did not establish full-trajectory identity. Subsequent paired tests checked exact nominal discrete endpoints and completion agreement within serialization roundoff (1.11×10^{-16}). These checks are reported separately rather than treated as interchangeable notions of replay.

C.3 Development severity selection

The same 30 nominally successful, brake-relevant development scenes are evaluated at every multiplier. The rule selects the largest tested multiplier with paired failure increase at least 0.10, a positive lower percentile-bootstrap endpoint, and strict-success fraction at least 0.50. Table 10 records both selected and unselected development settings.

Table 10: Development-only severity selection. All counts use the same 30 nominally qualified scenes. These are tuning results, not held-out estimates.

Command multiplier	Failures	Strict successes	Paired increase	95% bootstrap interval
1.00 (intact)	$0/30$	$30/30$	–	–
0.70 (selected)	$7/30$	$23/30$	23.3 pp	$[10.0, 40.0]$ pp
0.50	$9/30$	$20/30$	30.0 pp	$[13.3, 46.7]$ pp
0.35	$13/30$	$17/30$	43.3 pp	$[26.7, 60.0]$ pp

C.4 Monitor and fallback specification

The nominal-development fit uses 505 braking intervals and has one-step acceleration RMSE $2.206076\,\mathrm{m/s^2}$. Nominal braking intervals fit $a_{t+1} = ca_t + du_t$ by ordinary least squares without an intercept, using requested $u_t \le -0.25$. The reported conversion is

$$c = e^{-\Delta t/\tau}, \qquad d = (1-c)K, \quad K = 14.063216, \quad \tau = 0.080184\,\mathrm{s}. \tag{42}$$

The fit uses measured previous acceleration, whereas monitoring recursively propagates predicted acceleration. Its reported one-step RMSE is not a uniform recursive-rollout error bound.

At runtime the predicted acceleration is updated from the public commanded action. Positive observed-minus-predicted acceleration is scored only when predicted acceleration is at most $-0.20\,\mathrm{m/s^2}$. The instantaneous score divides this nonnegative underdeceleration residual by $0.50\,\mathrm{m/s^2}$; the EWMA uses weight 0.40 on the current score. Predicted acceleration and EWMA start at zero.

Table 11: Frozen nominal-development thresholds. They are quantiles of event-window maxima, not whole-episode false-alarm guarantees.

Score	$q = .95$	$q = .975$	$q = .99$
Instantaneous	10.579442	11.198487	11.909394
EWMA	7.465676	9.229865	10.506051

The nominal window includes score timestamps between the frozen event time and one second after it. Depending on the transition timestamp convention, its lower endpoint can include a score from the immediately preceding action interval. Threshold evaluation uses "greater than or equal to". Candidate monitoring runs throughout the episode, triggering once; after $0.5\,\mathrm{s}$ of simulation time, it latches a constant longitudinal command while preserving the controller's steering proposals. The service and emergency fallbacks request -3 and $-6\,\mathrm{m/s^2}$, mapped through K to approximately -0.213322 and -0.426645. The intervention still attenuates these negative commands. Neither requested value establishes achieved deceleration under degradation.

C.5 Complete clean-audit result

Table 12 reports the complete audit run. Here F_0, F_1 are physical failures among 24 nominal and 24 fault episodes, I_0, I_1 count episodes in which a fallback was actually released, and U is mean completion over all 48 condition rows. Alarm misses count fault episodes with no threshold crossing in the one-second calibration window after t_e (Appendix C.12); this is an alarm-budget metric, not a physical-deadline miss. All twelve rows fail the empirical screen of Equation (18).

Table 12: All frozen audit candidates. Fallback E/S denotes emergency/service. Every candidate is empirically inadmissible because $I_0 > 2$ out of 24; every candidate has $U \geq 0.85$. Without a wrapper, the same 24 audit pairs have $F_0 = 0$ and $F_1 = 5$.

Score	q/fallback	F_0	F_1	I_0	I_1	Alarm misses	U
EWMA	.950/E	1	6	6	8	$18/24$	.898060
EWMA	.950/S	1	8	6	8	$18/24$	.903506
EWMA	.975/E	1	6	6	6	$18/24$	.902498
EWMA	.975/S	1	7	6	6	$18/24$	.906848
EWMA	.990/E	0	5	3	3	$22/24$	.955081
EWMA	.990/S	0	6	3	3	$22/24$	.958021
Instantaneous	.950/E	2	6	7	8	$18/24$	.891311
Instantaneous	.950/S	2	8	7	8	$18/24$	.896979
Instantaneous	.975/E	1	6	6	8	$18/24$	.898087
Instantaneous	.975/S	1	8	6	8	$18/24$	.903477
Instantaneous	.990/E	0	5	3	5	$19/24$	.933246
Instantaneous	.990/S	0	6	3	5	$19/24$	.936751

The diagnostic selector maximizes common no-wrapper AUROC. The alarm-budget selector minimizes the recorded alarm-budget-miss rate, not a separately established physical-deadline miss. The empirical selector minimizes mean physical failure over the 48 rows. Ties use lower nominal intervention, then higher mean completion, then candidate ID. Because the admissible set is empty, none of these priorities selects an admitted policy.

The development and audit plots appear in Figure 1 in the main text.

C.6 Post-hoc calibration-window alignment diagnostic

Because calibration used one-second post-event maxima whereas candidate monitoring ran over the full episode, we performed a read-only diagnostic on the frozen logical execution plan. No threshold, candidate, physical outcome, or admission rule was changed. There were no pre-event first crossings, so the recorded first-trigger times determine exactly whether the first crossing lay inside the calibration window. Table 13 collapses the two fallback actions because they share the same monitor trigger.

Table 13: Nominal first crossings on the 24 audit scenes: whole-episode counting versus the one-second calibration window. The admissibility budget is at most 2/24 interventions.

Score	q	Whole episode	Calibration window
EWMA	.950	6/24	5/24
EWMA	.975	6/24	5/24
EWMA	.990	3/24	3/24
Instantaneous	.950	7/24	5/24
Instantaneous	.975	6/24	5/24
Instantaneous	.990	3/24	3/24

Window alignment reduces nominal first-crossing counts for lower quantiles and preserves the feasibility verdict at the closest operating points: both $q = .99$ monitors remain at $3/24 > 2/24$. On fault scenes, EWMA $q = .99$ changes from 3 to 2 first crossings while instantaneous $q = .99$ remains 5 to 5. These measurements quantify the calibration/operation mismatch while retaining the original Stage 5C decision. Appendix C.17 distinguishes this counting diagnostic from a closed-loop intervention evaluation.

Computational deduplication maps 576 logical candidate cells to 438 no-release baseline reuses and 138 intervention cells represented by 70 unique physical reruns. This changes computational work, not the candidate set or the number of independent scenes. The full equivalence assumptions are stated in Appendix A.9.

C.7 Post-hoc contextual-monitor sensitivity test

We tested whether the clean-audit rejection could be attributed primarily to the first-order response model. This analysis was designed after the Stage 5C audit and is therefore diagnostic rather than confirmatory. It uses only the retained 6000-series development traces and the frozen 7000-series Stage 5B no-wrapper traces; it does not inspect or execute the reserved 5000-series or either final 8000/9000 cohort.

A mandatory reproduction gate first reconstructs the original Stage 4 and Stage 5B monitor from raw traces. It reproduces the 505 nominal braking intervals, the original one-step RMSE $2.206076241206\,\mathrm{m/s^2}$, all six frozen nominal-development thresholds, and the stored Stage 5B scored audit trace row by row. Only after these checks pass is the contextual model evaluated. The predictor is fit on the same nominal braking domain, requested longitudinal action $u_t \le -0.25$, using previous measured acceleration, requested longitudinal action, previous speed, requested steering, absolute steering, and fixed quadratic/interaction terms. Non-intercept features are standardized within each training fold; ridge strength is selected by five seed-group folds and then refit on all 30 development scenes. The selected penalty is 1.0, with grouped-CV RMSE $1.389588\,\mathrm{m/s^2}$ and in-sample RMSE $1.250939\,\mathrm{m/s^2}$.

The predictor-to-score transformation is retained: positive observed-minus-predicted acceleration is scored only when predicted acceleration is at most $-0.20\,\mathrm{m/s^2}$, the instantaneous residual is divided by $0.50\,\mathrm{m/s^2}$, and the EWMA places weight 0.40 on the current score. Thresholds are quantiles of nominal one-second certification-window episode maxima. The offline contextual diagnostic counts strict threshold exceedances, whereas the executed Stage 5C candidate policies used a non-strict threshold comparison; the reported diagnostic counts are those produced by the archived implementation. Development discrimination becomes AUROC 0.621 for the instantaneous score and 0.632 for EWMA; the corresponding nominal/fault median maxima are 2.382/3.458 and 2.182/2.545.

The result separates prediction quality from operational admissibility. A read-only matched-CV replay on the same 505 intervals and five seed-group folds gives original/contextual mean-fold RMSE $2.230385/1.389588\,\mathrm{m/s^2}$ and pooled out-of-fold RMSE $2.240931/1.422398\,\mathrm{m/s^2}$; the contextual predictor is lower in all five folds. Because $\lambda = 1$ was selected on these development folds, this is a model-selection-stage diagnostic rather than an independent generalization estimate. Its score also gives measurable fault separation. Every tested operating point remains outside the nominal burden rule, with the strongest EWMA rows at three nominal triggers. The recorded outcome is no

Table 14: Post-hoc contextual-monitor audit on the frozen 24-pair Stage 5B no-wrapper cohort. Admission requires at most $2/24$ nominal interventions. No row is admissible.

Score	q	Threshold	Nominal triggers	Fault detections
Instantaneous	.950	5.243521	9/24	12/24
Instantaneous	.975	5.495156	6/24	9/24
Instantaneous	.990	5.589732	5/24	8/24
EWMA	.950	3.925778	3/24	10/24
EWMA	.975	4.064173	3/24	10/24
EWMA	.990	4.195099	3/24	8/24

frozen wrapper; the reserved 5000-series remains unopened. Appendix C.16 reports the fold-wise comparison and provenance, and Appendix C.17 states its scope.

C.8 Final no-wrapper campaign summary

Section 6.2 and Table 2 summarize the final no-wrapper cohorts; Figure 2 places their pooled result among the response policies. Scene-paired percentile-bootstrap intervals are $[8.3, 28.3]$ pp for held-out, $[8.3, 26.7]$ pp for replication, and $[10.8, 24.2]$ pp for the descriptive pooled result.

C.9 Post-hoc braking-demand fallback frontier

After completion of the frozen campaign, three separate descriptive extensions reused the same 60 held-out and 60 replication scenes under both nominal and $\beta = .70$ fault conditions. No scene selection, baseline outcome, fault severity, original campaign artifact, or trigger rule was changed across arms. Beginning at decision index 20, the PPO proposal was inspected. Three consecutive decisions with requested longitudinal command $u_t \leq -0.25$ and pre-action speed at least 20 km/h caused intervention on the third qualifying decision. PPO steering remained unchanged, and under the fault the frozen attenuation stayed downstream of the fallback command.

The arms vary only the response after this common trigger. *Full-brake latch* applies longitudinal command -1.0 to episode end. *Emergency latch* instead applies the already-declared Stage-5 emergency action $-6/K \approx -0.426645$ to episode end. *Emergency + release* applies that same emergency action until pre-action speed is at most 5 km/h for three consecutive decisions, then returns permanently to PPO control with no retrigger. The 5 km/h/three-decision rule was fixed before either new emergency arm was executed; it was not selected from their outcomes. Episode horizon and termination semantics were unchanged.

Each extension first ran four gate-disabled canaries, all of which reproduced their canonical baseline trajectories exactly. Each then produced 240 new-arm traces, all passing file-integrity, command-semantics, causal-trigger-order, and recorded-prefix checks. The common trigger fires in every episode, so the extensions isolate response consequence rather than detector selectivity. Because it applies online the same rule that defines the fault onset t_e (Appendix C), it fires two decisions after t_e in intact runs and cannot fire earlier in fault runs; it is aligned with the injected fault by construction. In the full-brake latch arm, pooled trigger timing is nearly overlapping across conditions: nominal median 126.5 decisions (mean 141.7) versus fault median 129.0 (mean 146.1). Table 15 gives the complete paired accounting.

The frontier is not monotone in a single scalar cost because authority and duration affect different consequences. Relative to no fallback, the permanent emergency latch repairs $18/21$ baseline fault failures but induces 23 new fault failures; the stronger full-brake latch raises repair to $21/21$ while also raising induced failures to 32. Bounding duration nearly removes the completion burden: pooled nominal arrival loss falls from 113 under the emergency latch to 10 and mean nominal completion change improves from -0.466 to -0.024. But repair falls to $8/21$, with 18 induced fault failures. Mean active intervention duration in the emergency + release arm is 20.16 decisions nominally and 24.24 under fault; 112/120 nominal and 108/120 fault episodes return to PPO.

For paired baseline and fallback outcomes with n_{rep} repaired and n_{ind} induced failures, the empirical failure-rate change is exactly $(n_{\mathrm{ind}} - n_{\mathrm{rep}})/n$. The pooled fault changes are $+4.2$ pp for emergency

Table 15: Post-hoc fallback frontier on the frozen final scenes. "Repaired" and "induced" are paired relative to the no-wrapper baseline within each condition. All arms intervene in every episode.

Arm	Cohort	Condition	Base fail	Arm fail	Repaired	Induced	Arrival lost	Δ completion
Emergency latch	Held-out	Nominal	0/60	12/60	0	12	56	−0.444
Emergency latch	Held-out	Fault	11/60	12/60	11	12	54	−0.414
Emergency latch	Replication	Nominal	0/60	13/60	0	13	57	−0.487
Emergency latch	Replication	Fault	10/60	14/60	7	11	52	−0.434
Emergency + release	Held-out	Nominal	0/60	15/60	0	15	4	−0.014
Emergency + release	Held-out	Fault	11/60	13/60	7	9	3	−0.006
Emergency + release	Replication	Nominal	0/60	19/60	0	19	6	−0.033
Emergency + release	Replication	Fault	10/60	18/60	1	9	9	−0.034
Full-brake latch	Held-out	Nominal	0/60	19/60	0	19	56	−0.457
Full-brake latch	Held-out	Fault	11/60	18/60	11	18	55	−0.441
Full-brake latch	Replication	Nominal	0/60	20/60	0	20	57	−0.501
Full-brake latch	Replication	Fault	10/60	14/60	10	14	54	−0.456

latch (repaired/induced $18/23$), $+8.3$ pp for emergency + release $(8/18)$, and $+9.2$ pp for full-brake latch $(21/32)$. The already-saved 20,000-resample descriptive paired-bootstrap 95% intervals are respectively $[-6.7, 14.2]$, $[0.0, 16.7]$, and $[-2.5, 20.8]$ pp; they are not confirmatory intervals for a randomized treatment effect. Repair rate alone is insufficient.

The recorded induced failures carry several types of adverse-event flag. Under nominal operation, emergency latch induces 25 failures with 24 vehicle-collision flags, 0 object-collision flags, and 11 road-departure flags; emergency + release induces 34 with 23 vehicle, 8 object, and 10 departure flags; full-brake latch induces 39 with 32 vehicle, 1 object, and 15 departure flags. Under brake-authority loss, the induced counts are respectively 23 (21 vehicle, 0 object, 7 departure), 18 (13 vehicle, 3 object, 9 departure), and 32 (31 vehicle, 0 object, 8 departure). These flag counts are not mutually exclusive: 5, 7, and 7 fault-induced episodes, respectively, carry multiple adverse flags. The logs do not identify collision geometry or assign causal responsibility, so we do not label these as rear-end events; the split only establishes that collateral failures include both collision and departure modes. Across these three policies, higher braking magnitude accompanies more repairs and more induced failures, while bounded duration preserves task utility with fewer repairs. All three retain the same universally firing trigger; this comparison measures complete response policies, not a causal decomposition of each mechanism. These arms evaluate the tested policies and do not establish unavoidable failure or impossibility of better fallback architectures.

The campaign archive retains the executed extension protocols and integrity records. The compact reproducibility package supplies selected numerical summaries and a separate file-transfer integrity manifest.

All three arms are explicitly post hoc because the 21 baseline fault failures were already known when the response family was designed.

C.10 Post-hoc compensation response study

After the braking-response frontier was known, we tested actuator-command compensation (archived under the name assimilation) on the same exposed final scenes. These extensions are explicitly post hoc and are not used to revise the frozen audit decision. The *oracle* arm activates only on faulted scenes at the known event time with gain $1/0.70$ before the frozen downstream attenuation; it yields 8/120 pooled fault failures, repairing 15 of the 21 baseline failures and inducing 2. The *fixed-gain* arm uses the common trigger with the same gain in every drive; it yields 16/120 fault failures (10 repaired, 5 induced) and 14/120 nominal failures. Commands remain clipped to the normalized actuator range, so requested commands below -0.70 cannot be exactly restored after 0.70 attenuation. The present archive does not attribute every residual oracle failure to clipping versus trigger/effect timing, and no such attribution is claimed here.

A validated family of twelve non-anticipative adaptive variants (C01–C12) varies only the number $k \in \{5, 7, 9\}$ of eligible observations, the predicted-braking eligibility threshold $\in \{-0.5, -1.0\}\,\mathrm{m/s^2}$ (column "Elig."), and the activation threshold $\widehat{\beta} \leq 0.85$ or 0.80, where $\widehat{\beta}$ is the median of the last k eligible observed-to-predicted acceleration ratios; the active gain is $1/\widehat{\beta}$,

clipped to $[1, 1/.70]$. Each run passes 240/240 observable-command, causal-trigger, file-integrity, and recorded-prefix checks. Table 16 reports pooled results. "Nom int." counts nominal scenes with any intervention; "Rep./ind." gives repaired/induced paired fault outcomes relative to no wrapper.

Table 16: Validated post-hoc adaptive compensation family on the exposed 8000/9000 scenes. The family was designed on already-exposed cohorts; rows are descriptive. None passes the post-hoc development carry-forward screen: at most 12/120 nominal interventions and at most 16/120 fault failures; Appendix C.13 gives its chronology and rationale.

Variant	k	Elig.	$\widehat{\beta}$ gate	Nom. int.	Nom. fail	Fault int.	Fault fail	Rep./ind.
C01	5	$-.5$	.85	63	1	81	21	3/3
C02	5	$-.5$	.80	57	2	76	21	3/3
C03	7	$-.5$	.85	49	5	75	21	2/2
C04	7	$-.5$	.80	34	4	66	21	2/2
C05	9	$-.5$	.85	37	2	67	19	2/0
C06	9	$-.5$	.80	23	0	51	20	1/0
C07	5	-1.0	.85	64	3	79	18	4/1
C08	5	-1.0	.80	58	3	74	17	4/0
C09	7	-1.0	.85	48	4	75	20	2/1
C10	7	-1.0	.80	34	2	66	20	2/1
C11	9	-1.0	.85	36	4	68	20	2/1
C12	9	-1.0	.80	25	2	54	18	3/0

The family records the relationship between response exposure and recovery. The oracle reference repairs 15 baseline failures and induces 2; the tested adaptive variants repair at most 4. That difference motivates further study of identification, timing, and response persistence. C06 records 23/120 nominal interventions and zero nominal failures, providing an example of the distinction between count burden and observed collateral cost. The original carry-forward screen remains unchanged. The table reports all twelve variants, with the exploratory interpretation consolidated in Appendix C.17.

C.11 Statistical interpretation and archived reference values

Within-cohort paired percentile bootstraps use 20,000 resamples and seeds 8628 and 8629. Cross-cohort comparison independently resamples scenes within each cohort with seed 8630. The pooled analysis resamples the 120 concatenated paired records with seed 8631 and is reported only as descriptive corroboration. It is not a balanced-stratum bootstrap or evidence of equivalence across the two generator subsets.

The recorded Wilson intervals are $[0.0, 6.0]\%$ for the held-out nominal rate and $[10.6, 29.9]\%$ for its fault rate. The nominal interval is not a field-safety bound: the tested population was selected to have no nominal failures. The held-out completion difference is -0.005077 with reported paired bootstrap interval $[-0.012946, -0.000088]$. These exploratory interval summaries depend on the eligible-scene sampling interpretation.

Why nominal selection must remain in the estimand. Consider an unselected population containing equal proportions of paired outcomes $(Y(1), Y(.7)) = (0, 1)$ and $(1, 0)$. Its unconditional average fault effect is zero. Selecting $Y(1) = 0$ retains only $(0, 1)$ and yields a conditional effect of one. This is not an error in estimating the selected effect; it is an error to call it the unconditional effect. The actual campaign also requires brake relevance and arrival, so its selection is more restrictive still.

McNemar values are preserved, not promoted. The archive's conventional two-sided calculation for induced/reversed counts b, c is

$$p_{\mathrm{ref}} = \min\left\{1,\ 2\sum_{j=0}^{\min(b,c)} \binom{b+c}{j} 2^{-(b+c)}\right\}. \tag{43}$$

It gives 0.0009765625 for $(11, 0)$, 0.001953125 for $(10, 0)$, and $9.536743164 \times 10^{-7}$ for $(21, 0)$. The standard symmetric-discordance null requires justification after conditioning on nominal success; it does not follow from paired execution or common random numbers. These values are therefore reported only as archived reference calculations rather than used to establish a randomized unconditional treatment effect. The within-scene intervention comparison and its explicitly conditional target remain the substantive evidence.

The held-out failure breakdown comprises ten collision episodes and two road-departure episodes, with one overlap. The onset-to-first-adverse interval has median $10.40\,\mathrm{s}$ and interquartile range $[8.15, 15.00]\,\mathrm{s}$.

Finally, the first-adverse timestamp uses the start index of the transition in which an adverse flag was observed; its interval has control-step resolution and is not a continuous impact time. No certified physical safe-release frontier is claimed; the bounded-release arm is a descriptive policy point, not a safety certificate.

C.12 Empirical separation of the recorded alarm channel

Let $Z = \mathbf{1}\{\text{score crosses its fixed threshold within the recorded event window}\}$. The TV distance between two Bernoulli laws is the absolute difference of their success probabilities. Applied to Table 14, the contextual EWMA yields:

q	Nominal alarm	Fault alarm	$\widehat{\mathrm{TV}}_Z$	False-trigger + miss rate
.95, .975	$3/24$	$10/24$	$7/24$	$17/24$
.99	$3/24$	$8/24$	$5/24$	$19/24$

These are empirical alarm-channel quantities, algebraically determined by the recorded counts. At population level data processing gives $\mathrm{TV}(\mathcal{L}_0(Z), \mathcal{L}_1(Z)) \le \mathrm{TV}(Q_0, Q_1)$ when Z is measurable from that same history. The empirical difference is not, by itself, a population confidence bound. More importantly, even a justified *lower* bound on history TV cannot prove an upper limit on detectability or be substituted for TV in the decreasing right side of Equation (14). For example, retaining detection $8/24$ while targeting nominal burden $2/24$ would require history separation at least $6/24$; a witness of $5/24$ does not exclude it.

The scores are timestamped at `time_s` + `dt_s` and counted in $[t_e, t_e + 1\,\mathrm{s}]$, with t_e determined by the intact reference trajectory. This is evaluator-side reference alignment for the archived alarm-budget window, not an online estimate of the last-safe physical cutoff. Applying these alarm-channel quantities to an operational history law requires that the time-origin information used by the statistic be available under that declared observation protocol. The diagnostic reports what was observed in that channel and does not claim that the full-history frontier numerically predicted the audit rejection.

C.13 Analysis lineage and carry-forward rule

These experiments were developed after the original no-wrapper outcomes were available. First, the three braking-response arms were evaluated. Oracle and fixed-gain compensation followed, yielding repaired/induced fault counts $15/2$ and $10/5$. The first online adaptive arm then yielded $24/120$ fault failures (4 repaired, 7 induced), $8/120$ nominal failures, and $89/120$ nominal interventions. It is a distinct policy, not one of C01–C12; retaining its result prevents a claim that every tested compensation implementation was non-harmful.

The subsequent development protocol fixed the 12 variant settings before their correctly instantiated sweep: $k \in \{5, 7, 9\}$, eligibility acceleration in $\{-.5, -1.0\}$, activation ratio in $\{.85, .80\}$, two consecutive qualifying estimates to activate, three consecutive estimates at ratio $\ge .95$ to release, and gain ceiling $1/.70$. The carry-forward screen required at most 12 nominal interventions and at most 16 fault failures in 120 scenes. The latter is intentionally the *already observed* fixed-gain result: adaptation was required to retain that response's observed recovery while reducing exposure. The threshold is an exploratory engineering benchmark, not an independently calibrated safety guarantee.

Packaging and replay-validation failures occurred in earlier tuning harness versions. Inspection showed that those workers instantiated the support module's adaptive controller rather than the candidate-specific controller. Their outputs were excluded as executions of the wrong policies, not because their outcomes were unfavorable. The corrected V3 worker imports the candidate controller, runs in a separate result root, and produced all 12 rows reported here. All completed rows, including those failing the carry-forward rule, are retained. The original failed runs were preserved rather than reclassified. The retained protocol is reproduced in `reproducibility/protocols/ADAPTIVE_TUNING_V2_PROTOCOL.md`; its archived identifier documents a local frozen specification, not public preregistration. No candidate passed, and no new test cohort is represented as evaluated in this paper.

C.14 Evidence budget of the frozen adaptive estimator

A read-only diagnostic replays the already-executed C01–C12 estimator on the exposed 8000/9000 no-wrapper traces; it introduces no new estimator, changes no policy, and does not open the reserved 5000-series. Before first activation the adaptive gain remains one, so the estimator sees the same command/kinematic prefix as the no-wrapper trajectory. Replay therefore identifies the first activation exactly up to that point. As an implementation check, the replay reproduces all 2,880 archived first-trigger outcomes (12 candidates $\times$ 240 scene-condition executions) with zero mismatches.

The diagnostic fixes "fresh fault-exposed evidence" as eligible ratio updates strictly after the frozen fault onset: the update timestamped at the onset still reflects the preceding pre-fault command. Under the most permissive eligibility rule, predicted acceleration $\leq -0.5\,\mathrm{m/s^2}$, only 1/21 baseline fault failures accumulates fewer than five fresh eligible observations before its first adverse transition. Thus, relative to the first adverse transition, scarcity of eligible observations is not the dominant explanation for this estimator family's misses; an earlier effective-response deadline could impose a stricter evidence budget. The ratio itself is informative: using the median of the first five fresh eligible observations, fault versus intact operation gives AUROC .645 with paired-scene bootstrap 95% interval $[.621, .676]$ on 118 paired-complete scenes. Lower ratios are predeclared as more fault-like; the score direction is not reversed after inspection.

Table 17: Read-only evidence-budget replay for all twelve adaptive variants C01–C12. "Fail pre" is activation at or before fault onset; "fail post" is activation after onset but before the first adverse transition. The last two columns show any activation in the 99 non-failing fault runs and 120 intact runs. Failure columns are censored at the first adverse transition; non-failing fault and intact activation counts use the full episode, so the columns are not equal-window detection-rate comparisons.

Variant	k	Elig.	$\widehat{\beta}$ gate	Fail pre	Fail post	Fault nonfail	Intact
C01	5	$-.5$	.85	1	8	71/99	63/120
C02	5	$-.5$	.80	1	8	66/99	57/120
C03	7	$-.5$	.85	0	9	65/99	49/120
C04	7	$-.5$	.80	0	8	56/99	34/120
C05	9	$-.5$	.85	0	7	57/99	37/120
C06	9	$-.5$	.80	0	6	41/99	23/120
C07	5	-1.0	.85	1	8	69/99	64/120
C08	5	-1.0	.80	1	8	64/99	58/120
C09	7	-1.0	.85	0	9	65/99	48/120
C10	7	-1.0	.80	0	8	56/99	34/120
C11	9	-1.0	.85	0	7	58/99	36/120
C12	9	-1.0	.80	0	6	44/99	25/120

Table 18: Fault information in the frozen acceleration-ratio statistic. Each row uses the median of the first k fresh eligible observations. AUROC uses the predeclared direction "smaller ratio is more fault-like"; intervals are 20,000-resample paired-scene bootstraps.

Elig.	k	Paired	Median intact	Median fault	AUROC [95%]
$-.5$	1	120	1.500	1.232	.663 $[.632, .698]$
$-.5$	3	120	1.071	.907	.670 $[.647, .700]$
$-.5$	5	118	1.038	.907	.645 $[.621, .676]$
-1.0	1	120	1.500	1.232	.663 $[.633, .698]$
-1.0	3	120	1.071	.907	.670 $[.646, .700]$
-1.0	5	118	1.038	.907	.645 $[.621, .677]$

The positive conclusion is therefore signal-level rather than deployment-level: the injected 30% actuator loss leaves measurable information in the tested online statistic. The unresolved step is converting that partially overlapping evidence into selective and timely action. More permissive settings activate before adversity in more failing runs, but they also activate frequently in intact operation; more conservative settings lower intact activation while leaving more failures untriggered. This post-hoc diagnostic explains the observed selectivity–timeliness tradeoff of the executed estimator family; it does not establish an impossibility result for stronger fault-diagnosis methods.

C.15 Matched failure-case trajectory diagnostic

A read-only diagnostic joined all 21 final-cohort no-wrapper fault failures to their nominal and fault scalar traces, verified the referenced files against the retained archive, and aligned the pre-action step of the first adverse transition. There are 11 held-out and 10 replication failures. Nominal trace data are available at the adverse transition for all 21 cases, so no extrapolation beyond a terminated nominal trajectory is used.

Using a numerical divergence tolerance of 10^{-7}, fault and nominal speed first differ one control step after the frozen fault onset in all 21 cases; this is primarily a propagation check because onset is defined at a braking command whose applied value is attenuated. Requested steering first differs one step after onset in 20/21 cases and after 12 steps in the remaining case. In the ten-step pre-adverse window, the median steering RMS difference is 0.316 normalized command units and the median maximum absolute difference is 0.550.

At the first adverse transition, the fault trajectory is faster than nominal in 11 cases and slower in 10; the median fault-minus-nominal speed difference is $+0.137\,\mathrm{m/s}$, with IQR $[-0.615, 1.079]$ and range $[-2.464, 4.160]$. The 21 episodes comprise 18 with at least one collision flag and 5 with a road-departure flag, with two collision/departure overlaps. These measurements are consistent with propagation through the closed-loop controller, but they do not identify the mediation path, a common stopping-distance mechanism, collision geometry, or causal responsibility for individual events. Late comparisons are aligned by simulation time rather than route position, so spatial phase differences can contribute to speed and steering differences.

To test whether simple early propagation distinguishes eventual failure, we fixed three onset-anchored statistics before the full comparison: nominal braking depth over the first 2 s, fault-minus-nominal speed at $+1$ s, and steering-command RMS divergence over the first 2.5 s. Fault-dependent speed and steering measurements are censored at the first adverse transition; nominal braking depth uses the intact reference trajectory. On the exposed final cohorts, braking depth has AUROC $.447$ with bootstrap 95% interval $[.319, .578]$ (21 failures, 99 non-failures); $+1$ s speed excess has AUROC $.552$ with interval $[.408, .692]$ (21 failures, 94 non-failures with an available value; the other five lack a paired trace sample at $t_e + 10$, so the extraction rule leaves them missing without extrapolation); and steering RMS has AUROC $.493$ with interval $[.366, .621]$ (21 versus 99). All intervals include $.5$. We therefore use these quantities as propagation diagnostics rather than discriminative mechanism evidence. The analysis is exploratory because the cohorts were already exposed; no metric direction was reversed after inspection.

TTC at onset remains uncomputed because the scalar traces do not contain the documented object-relative positions, velocities, and geometry required for a collision-course definition. Likewise, attributing residual oracle failures to clipping versus delay requires intervention or ablation evidence; observing saturation alone would not prove causation. The reproducibility package contains the scripts, matched-CV prediction records, 21-case table, and complete 120-scene onset-control metric table with its analysis specification and output report. The packaged rows reproduce the onset-control AUROCs and frozen bootstrap intervals; regenerating the metrics from trajectories still requires the separate raw paired traces.

C.16 Matched development-CV result

The read-only replay refits the original no-intercept OLS predictor $a_{t+1} = ca_t + du_t$ and the contextual ridge predictor on the same 505 nominal intervals ($u \leq -.25$), using the same five seed-group folds and training-fold standardization. It first reproduces the original full-development

fit RMSE $2.206076241206\,\mathrm{m/s^2}$ and the contextual selection result $\lambda = 1$, mean-fold RMSE $1.389588484841\,\mathrm{m/s^2}$.

Table 19: Matched five-fold development diagnostic on the same 505 braking intervals. The contextual penalty was selected on these folds, so these are model-selection-stage diagnostics rather than an independent or nested generalization estimate.

Fold	Original OLS RMSE	Contextual $\lambda = 1$ RMSE
0	2.2860	1.3599
1	2.5759	1.9710
2	2.3035	1.0395
3	2.2513	1.2043
4	1.7352	1.3733
Mean of fold RMSEs	2.2304	1.3896
Pooled out-of-fold RMSE	2.2409	1.4224

The contextual predictor is lower in all five folds. Relative to the original model, the mean-fold RMSE decreases by 37.7% and pooled out-of-fold RMSE by 36.5%. The comparison uses only the 6000-series development traces; it does not read the 7000/8000/9000 traces for fitting, change a threshold, or alter the completed audit. The complete JSON metrics and out-of-fold rows are archived with the diagnostic outputs.

C.17 Limitations and evidence scope

Scope of the formal guarantees. The results apply to the stated task, action family, loss, completion class, and information channel. The composed theorem is stated for one commitment with a common pre-response history law and conditioned validity; Appendix A.14 gives a conditional stagewise extension under additional stage-specific premises. Covered risk is conditional Bayes risk, and the risk-set and testing results use classical mathematics. The contribution is their compatible specification for physical assurance. The synthetic completion class and law-cell partition are supplied externally. A class can omit a relevant mechanism, and auditing its listed members does not establish universal coverage. Marginal risk-report bias is distinct from conditional calibration, and the value-of-information result is an expectation statement. General-space extensions require specified regular conditional distributions.

From scores to physical risk. The δ-validity contract requires evidence about action-conditioned consequence loss. Acceleration RMSE alone does not establish a collision-risk tolerance. On matched five-fold seed-group splits of the 505 development intervals, original/contextual mean-fold RMSE is $2.230/1.390\,\mathrm{m/s^2}$ and pooled out-of-fold RMSE is $2.241/1.422\,\mathrm{m/s^2}$; the contextual penalty was selected on those folds, so this is a development diagnostic rather than an independent generalization estimate. Neither score AUROC nor empirical binary alarm TV upper-bounds full-history separation. The physical-loss lower bound requires its missed-response consequence premise and an appropriate full-history TV bound. MetaDrive does not measure the composed certificate or a last-safe response time. The known command multiplier β is not the loss-unit actionwise tolerance δ.

Screening and statistical admission. The 24-scene audit implements a frozen empirical burden/progress screen. It is not a population safety certificate. For a fixed Bernoulli intervention indicator, even $0/24$ has a one-sided exact 95% upper bound $1 - 0.05^{1/24} \approx 0.1173$, above a 10% budget. The observed $3/24$ versus $2/24$ distinction therefore determines the executed screen without establishing a population-level separation. Appendix A.7's simultaneous finite-candidate confidence gate is analytical and was not executed. Window alignment preserves the narrow empirical rejection result; it does not evaluate a new closed-loop window-limited policy. Response-sensitive cost budgets are a prospective extension, not a replacement for the completed count gate.

Exploratory response comparisons and selection. The braking, compensation, and contextual-model analyses were developed after earlier outcomes had been examined. The adaptive compensation family uses the already-exposed 8000/9000 cohorts. Its estimates are descriptive, subject to

multiplicity and selection bias, and establish neither confirmatory efficacy nor deployment admission. The oracle arm has privileged timing and fault knowledge and is a reference intervention, not a proved optimal ceiling. Appendix C.13 reports the earlier adaptive arm and the later candidate family separately; their counts do not support class-wide non-harm. Reproduction checks for existing scores establish numerical consistency for that implementation, not independent validation of the new monitor's full causal deployment design.

Population, endpoints, and mechanisms. The campaign conditions on nominal success and braking relevance, so its nominal zeros and paired effects describe that selected population. One controller checkpoint, one simulator configuration, and one command-attenuation fault family define the empirical scope. The released actor is evaluated separately from AIM's interactive-learning algorithm. In the 21 final-cohort failures, the adverse-transition speed is above the time-aligned nominal value in 11 cases and below it in 10, so the failures are not uniformly characterized by excess ego speed at that instant. These late comparisons are time-aligned rather than route-position aligned. A pre-fixed onset-control comparison against non-failing fault episodes found no clear discrimination from initial braking depth, $+1$ s speed excess, or early steering divergence; its AUROC intervals all include .5. Collision/departure flags do not establish geometry, injury severity, mediation, or causal responsibility. The recorded onset-to-adverse interval is not a certified response deadline. Command clipping limits exact reciprocal compensation below -0.70; residual oracle failures are not causally attributed among saturation, timing, and other closed-loop effects.

Evidence provenance and next validation. File-integrity, semantic, and replay checks address specified implementation properties; none establishes unrestricted correctness. Experimental reproduction of the full MetaDrive campaign requires the checkpoint, trajectories, execution code, and matching simulator assets in addition to the manuscript bundle. Appendix C.13 specifies which tuning runs were excluded and why.

Future work. The next validation should freeze the response policy, consequence contract, nominal response-cost function, and statistical admission rule before a fresh campaign. Fresh calibration and test cohorts should span additional controllers, scenario distributions, actuator and sensing faults, and operating regimes. Route-position and object-relative geometry should support collision-course and last-safe-time labels rather than time-index surrogates. The most important empirical target is prospective evaluation of the composed upper certificate: externally auditable coverage and response-success labels, a declared miss-rate bound, and a response-loss bound measured on held-out scenes. Sequential extensions should then be evaluated under repeated interventions using the stagewise conditions of Appendix A.14.

D EXPANDED RELATED WORK

Decision risk and identification. POMDPs and imperfect-information games study decisions under hidden state (Kaelbling et al., 1998; Reif, 1984). The finite covered-model quantity ρ is conditional Bayes risk. Partial identification separates what observational laws determine from what a declared class of completions supplies (Manski, 2003; Pearl, 2013). The exact and approximate reporting limits use that perspective: a narrow identified set describes precision, and its upper endpoint describes the largest risk compatible with the assumptions. Hoeffding concentration controls sampling error within a specified class (Hoeffding, 1963); it does not identify the class itself. The contribution here is the common consequence specification that allows these components to enter the composed commitment bound.

Quickest change detection. Page's cumulative-sum formulation, Lorden's delay criterion, and the sequential-detection literature formalize the trade between false alarms and delayed detection (Page, 1954; Lorden, 1971; Tartakovsky et al., 2014). The EWMA residual is a classical monitoring construction, not a new detection algorithm. Our history-law frontier fixes the information available before a response can become effective and introduces an explicit missed-response consequence premise. This differs from equating a statistical stopping-time objective with physical protection. The framework can accept a CUSUM, likelihood-ratio, or calibrated learned monitor when its observation and execution assumptions are specified. OOD learnability and conditional inference results

likewise emphasize the restrictions required for distributional guarantees (Fang et al., 2022; Barber et al., 2021).

Fault detection, isolation, and accommodation. Model-based diagnosis develops residual generators that distinguish faults from uncertainty, disturbances, and ordinary plant behavior (Chen & Patton, 1999; Isermann, 2006). Fault-tolerant control connects diagnosis with reconfiguration and accommodation under residual control authority (Blanke et al., 2016). Thus residual-based monitoring and reciprocal actuator compensation are established techniques. The MetaDrive response study uses simple instances to expose the difference between a command-level fault, its closed-loop effect, and the outcome of a proposed remedy. It measures that distinction with paired repair, induction, completion, and intervention counts. It does not claim that adaptive compensation itself is new or that a tuned candidate family characterizes every fault-tolerant controller.

Safe distance, runtime assurance, and learned safety filters. Responsibility-Sensitive Safety includes safe-distance and proper-response reasoning under explicit kinematic assumptions (Shalev-Shwartz et al., 2017). Equation (1) is an elementary stationary-obstacle stopping condition, not the full RSS model. Simplex, barrier functions, reachability, shielding, and predictive safety filters construct protective responses under their respective model and feasibility assumptions (Seto et al., 1998; Sha, 2001; Ames et al., 2017; Bansal et al., 2017; Alshiekh et al., 2018; Wabersich & Zeilinger, 2023). These methods motivate retaining entry conditions, disturbance bounds, and available actuation time as separate assurance evidence. Our composed statement expresses the risk contribution of continuation and response on one horizon, with latent-stratum validity stated explicitly.

Conformal monitoring and model adaptation. Conformal prediction for STL runtime verification calibrates uncertainty either in trajectory predictions or in the specification's satisfaction measure (Lindemann et al., 2023). Robust variants address distribution shift under additional divergence assumptions (Zhao et al., 2024). SODA-MPC couples a conformalized OOD signal to reachability-based fallback under a specified pedestrian-motion bound (Contreras et al., 2025). These precedents already join predictive uncertainty with runtime response. Our question is how their evidence obligations compose when the consequence contract or the response itself is uncertain. Predictive coverage, response feasibility, and low closed-loop loss require distinct supporting premises; the audit records which are established for each implementation.

Standards and safety-case context. ISO 21448:2022 addresses hazards from functional insufficiencies within the safety of intended functionality and explicitly excludes faults covered by ISO 26262 (International Organization for Standardization, 2022). The command-attenuation experiment is therefore not presented as a SOTIF compliance test or as a direct instance of every SOTIF category. It is closer to the fault-handling timing of ISO 26262 (International Organization for Standardization, 2018), whose fault detection, fault reaction, and fault-tolerant time intervals play roles similar to our decision cutoff t_d, execution-delay bound $\bar{L}$, and contract deadline t_{cert}; we use this only as an analogy, not as a mapping of its requirements. UL 4600 organizes safety arguments and evidence for autonomous products (UL Standards & Engagement, 2023); guaranteed-safe-AI frameworks similarly organize models, specifications, and verification (Dalrymple et al., 2024). Our risk sets and escape terms provide mathematical vocabulary for selected evidence gaps, rather than a one-to-one translation of standards categories. No standards certification is asserted. Appendix C.17 consolidates the empirical and formal scope.

E Experimental artifacts and reproducibility

E.1 Programs and numerical records

The reproducibility package contains the manuscript source, bibliography, figures, numerical claim checks, and analysis utilities used for the reported tables and diagnostics. Historical experiment filenames are retained inside the package where changing them would break the lineage between a program and its archived output, but the paper does not rely on those names as scientific terminology.

The external synthetic archive holds the programs and result files for the blind-fork, triple-conflict, crosswalk, completion-class, and mechanism-holdout studies; the compact package includes their

SHA-256 digests and the rational-arithmetic check of Appendix B.4. Appendix B specifies their laws, endpoints, and sampling procedures. The external campaign archive records the fixed controller identity, simulator configuration, qualification rules, candidate catalog, audit table, no-wrapper outcomes, response-family summaries, and the analysis chronology used in Appendix C. The reported campaign uses 30 development scenes, 24 audit scenes, and two disjoint 60-scene final cohorts.

Post-freeze diagnostics. The matched development-CV, 21-case trajectory comparison, and 21-versus-nonfailure onset-control analysis are read-only analyses of retained data. They do not run MetaDrive or change a policy, threshold, cohort, or outcome. The matched CV uses only the 6000-series development traces. The trajectory and onset-control diagnostics use the already exposed 8000/9000 no-wrapper cohorts and are labeled accordingly. Their scripts, retained matched-CV and case-table records, complete onset-control metric table and output report, and run instructions are included in the reproducibility package. The packaged prediction and case rows reproduce the matched-CV aggregates and failure-case summaries. The 120-scene onset table reproduces all three AUROCs and their frozen 20,000-resample bootstrap intervals. Refitting the predictors or regenerating trajectory metrics requires the separate raw inputs.

E.2 Experiment versioning and verification

An earlier pilot campaign with 65 development scenes, $\beta = .50$, and three held-out wrapper arms is superseded by the campaign reported here, whose development rule selected $\beta = .70$. The two are not pooled; the pilot's recovery-window and wrapper-effect estimates are not evidence for the current campaign.

The clean audit returned `NO_FEASIBLE_CANDIDATE`; neither final cohort therefore includes an admitted frozen wrapper. The final experiments establish the conditional no-wrapper stress effect and its disjoint-seed replication. Later contextual-monitor, braking-response, compensation, and trajectory diagnostics reuse retained data and are reported separately from the frozen audit. Appendix C.13 records the analysis order, the earlier adaptive arm, the later twelve-variant family, and the reason invalid harness executions were excluded.

Package contents and reproducibility boundary. The anonymous source bundle builds the paper directly. The accompanying reproducibility bundle contains analysis scripts, selected result tables/JSON records, environment notes, and integrity metadata for file transfer. Raw MetaDrive trajectories and the controller checkpoint are not embedded in this compact package; a complete end-to-end rerun additionally requires those archived data, the simulator assets, and the recorded software environment. The package therefore supports manuscript rebuilding and numerical/analysis replay from the retained records, while the full campaign archive is the source for complete simulator reproduction.