CASTLE: CAUSALLY ATTRIBUTED SOFT CONJUNCTIONS FOR SOUND AND INFORMATIVE SAFETY CERTIFICATES

Anonymous authors
Paper under double-blind review

ABSTRACT

A multi-predicate safety specification is almost always relaxed into a differentiable conjunction, most often by the log-sum-exp conorm. The relaxation is then used as a *penalty*. Turning it into a *certificate* — a state predicate a runtime monitor may act on — additionally requires choosing the value at which the conjunction is declared satisfied, and the standard choice, $\log K/\beta$, is not the tightest admissible one. We make three contributions. **(i)** For the weighted conorm we derive the minimal admissible robustness threshold $\tau_\omega = \log(1/\omega_{k^\star})/\beta$, pinned to the budget mass on the binding predicate; the resulting certificate is sound for *every* budget and carries no slack beyond the conorm's own gap, whereas the uniform threshold of prior work adds $\log(K\omega_{k^\star})/\beta$ of pure conservatism. **(ii)** We show that the quantity that should decide where budget goes is not bindingness but *control authority*, and introduce IARS, an interventional attribution that measures per-predicate action sensitivity conditional on the causal order. It is identically zero for agent-independent monitors and for restatements of an already-binding predicate. **(iii)** We prove that the resulting decomposition — a soft conjunction over attributable predicates conjoined with an *exact* nominal check on the rest — is sound, and we formalise the design principle that robustness margin should be spent only where the agent can act. Across three environments and three seeds, CASTLE raises the activation rate of a *sound* certificate from 0.215 to 0.381 on a specification containing a non-attributable predicate — at zero certified-but-unsafe steps — while matching or improving the safety–performance frontier; removing the exact monitor reproduces 3253 certified-but-unsafe steps, while the product t-norm certifies 1463 such steps at its natural threshold and the Łukasiewicz certificate never activates at all.

1 INTRODUCTION

Constrained reinforcement learning represents a safety specification as a conjunction of predicates and then relaxes the conjunction, because the exact conjunction $\min_k m_k$ has unusable gradients. The log-sum-exp (LSE) conorm $C_\beta(m) = -\frac{1}{\beta}\log\sum_k e^{-\beta m_k}$ on margins $m_k = -\psi_k$ is the relaxation of choice in recent neuro-symbolic safe-exploration work: it is dense, analytic and sandwiched by $\min_k m_k$ and $\min_k m_k - \log K/\beta$ (Anonymous, 2027). The relaxation is used in two roles. As a *penalty* it feeds a Lagrangian dual. As a *certificate* it supports a state predicate — "the conjunction holds with slack $\rho + \delta$" — which a runtime monitor can act on (Anderson et al., 2020; Fulton and Platzer, 2018; Hamel-De le Court et al., 2025).

The two roles impose different requirements, and it is the second that we find under-served. A certificate is useful only if it fires often enough to be informative: a predicate that is almost never true carries no information for action selection, for hand-over decisions, or for shielding (Hamel-De le Court et al., 2025; Rabiee and Safari, 2025). Yet the threshold inherited from the sandwich, $\log K/\beta$, is the tightest one admissible *when the conjunction is unweighted*. Weighted conorms — which arise as soon as one admits that predicates are not interchangeable — admit a strictly smaller threshold, and the smallest admissible value depends on the budget mass placed on the predicate

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.

that currently binds. Section 4.1 derives it: $\tau_\omega = \log(1/\omega_{k^\star})/\beta$, with $k^\star$ the binding predicate. The certificate built on it is sound for every budget, and the uniform threshold is exactly the $\omega \equiv 1/K$ special case, i.e. the most conservative member of the family. On our attribution benchmark this is not a cosmetic difference: the LSE certificate with the uniform threshold fires on 0.215 of steps, against 0.381 for the weighted one.

That immediately raises the question the second half of the paper is about: *which* predicates should receive budget? The answer used implicitly by the conorm — and optimised explicitly by the temperature controller of Anonymous (2027) — is bindingness: weight follows $\mathrm{softmax}(-\beta m)$. Bindingness is an associational quantity. A safety monitor whose residual the agent cannot influence at all (a co-worker's trajectory, a wind gust, a sensor channel) will be violated often, hence will be small, hence will attract almost all the weight, and the conjunction will be dominated by a constraint that no action can satisfy. Redundant restatements of an already-binding predicate are a second, quieter failure mode: they split the budget without adding control authority. Figure 2 (Appendix A) shows both effects on a specification with eight predicates: 82% of the associational budget lands on the agent-independent monitor.

We replace associational weighting with an interventional one. IARS (interventional action-residualised sensitivity) computes, per predicate, the norm of the action Jacobian of its residual relative to the population, and residualises it against the predicates that are *more binding* — the causal order of the conjunction. The result is a gate that is identically zero for agent-independent monitors (their Jacobian vanishes) and for co-oriented restatements (their Jacobian is already spanned), and positive for constraints that carry genuine, non-redundant authority. Because the gate is built from derivatives, it is a property of the specification-and-state, not of the residual magnitudes, and therefore immune to the confounding that misleads softmax weighting.

The gate does more than clean up the penalty. It propagates into the certificate through the threshold: budget moves off non-attributable predicates, $\omega_{k^\star}$ rises, τ_ω falls, and the certificate activates more often. It also licenses a change in *semantics* that we consider the main conceptual point. Robustness margin is a resource: it is what lets a certificate survive model error and discretisation. Spending it on a constraint the agent cannot hold buys nothing, and costs activation rate. We therefore split the specification into an attributable part, certified softly with margin $\rho + \delta$, and a non-attributable part, certified by an *exact* nominal check $\psi_k \le -\delta_m$. The soft relaxation is retained exactly where a gradient is useful; where no action can help, a hard check suffices. Theorem 2 shows the two together remain a sound certificate of the full specification.

Contributions.

- **Exact thresholds for weighted soft conjunctions** (Section 4.1). We give the tight sandwich for the weighted conorm (Lemma 1) and prove that $\tau_\omega = \log(1/\omega_{k^\star})/\beta$ is not merely sufficient but *minimal* (Corollary 1). Soundness of the resulting certificate holds for arbitrary budgets (Theorem 1), and activation rate is governed by τ_ω and converges to the maximum achievable rate as β grows (Proposition 1).
- **Interventional attribution for conjunctions** (Section 4.2). We define IARS, prove its two invariances (Lemma 2), and explain why residualisation must be taken against the causal order rather than against the full action span: in a d-dimensional action space the latter saturates and becomes uninformative, which we verify empirically.
- **A decomposition theorem and a design principle** (Section 4.3). The attributable/exact-monitor split is sound (Theorem 2), its activation rate dominates the monolithic alternative (Proposition 1), and it converts a certificate that empirically never fires into one that fires often with no loss of soundness.
- **Experiments** (Section 5). Three vectorised environments, eight methods, five ablations, three seeds, plus a certificate audit that counts certified-but-unsafe steps. We find that the product and Łukasiewicz t-norms, evaluated at their natural threshold, certify 1463 steps, while the LSE family certifies none; that t-norm certificates can also fail by being vacuous, as the Łukasiewicz one does (it never activates); that the uniform-threshold certificate is nearly vacuous on a specification with a non-attributable predicate; and that CASTLE dominates the safety–performance frontier on the benchmark where attribution matters.

2 RELATED WORK

Constrained and safe RL. Constrained Markov decision processes (Altman, 1999) and their primal-dual solution (Achiam et al., 2017; Tessler et al., 2019; Chow et al., 2017) remain the workhorse of safe RL; the PPO-Lagrangian variant we build on is standard practice (Schulman et al., 2017). Safety layers project proposed actions onto a feasible set (Dalal et al., 2018), and shielding restricts the action space using a verified model (Hamel-De le Court et al., 2025; Rabiee and Safari, 2025). Certificates based on control barrier functions give per-step guarantees (Ames et al., 2019; Yang et al., 2025) and have recently been composed with soft-min approximations of multi-constraint conjunctions (Wang et al., 2025; Safari and Hoagg, 2025). Our work is complementary: we do not propose a new certificate class but fix the threshold of the smooth-conjunction certificate, and we address which predicates may enter it. Benchmarks (Ray et al., 2019; Ji et al., 2023) and surveys (Kushwaha et al., 2025) structure the field; the recent shift towards vision-language-action policies brings the same questions into embodied settings (Cao et al., 2025b; Tang et al., 2026; Hu et al., 2025a; Gu et al., 2025; Ma et al., 2026).

Neuro-symbolic conjunctions. Differentiable logic replaces Boolean connectives by t-norms and t-conorms, of which the Gödel $(\min)$, product, Łukasiewicz and LSE families are the standard choices (Badreddine et al., 2022). The LSE conorm is the smooth minimum of choice whenever gradients must remain dense (Agazzi et al., 2025; Safari and Hoagg, 2025). Neuro-symbolic RL adds logical structure to policies and value functions (Delfosse et al., 2023; Shindo et al., 2025; Pfanschilling et al., 2025; Shindo et al., 2024; Skryagin et al., 2024), and verified exploration lifts symbolic policies into neural ones under a safety argument (Anderson et al., 2020; Fulton and Platzer, 2018). Our work keeps the operator and its differentiability, but treats the pair (budget, threshold) as the object of study.

Causality and credit assignment. The causal machinery of Pearl (2009) has been used in RL to prioritise causally relevant features and actions (Cao et al., 2025a), to redistribute credit across time (Mesnard et al., 2021), and to attribute behaviour to data (Hu et al., 2025b). These methods target reward-relevant variables. Ours targets the constraint side: which predicates of a conjunction possess independent control authority, which is what decides whether a certificate is meaningful.

3 PRELIMINARIES

Specifications and residuals. A safety specification is a set of K predicates. Predicate k exposes a residual $\psi_k(s,a) \in \mathbb{R}$ with $\psi_k \le 0$ meaning *satisfied*, and the specification holds iff the *worst* residual obeys $\max_k \psi_k \le 0$. We write the margin $m_k = -\psi_k$, so $m_k \ge 0$ is safety, and the conjunction of the specification is the pointwise minimum of the margins, $\min_k m_k = -\max_k \psi_k$. The monitor reports $\hat{\psi}_k(s)$; we assume $\hat{\psi} = \psi$ for the benchmark environments and treat corrupted reports as a design axis for the attribution rather than as noise to be filtered. Following **?**, the margin used by the conjunction is *endogenous*:

$$m_k^\theta(s) = \hat{m}_k(s) + \rho \tanh \varepsilon_k^\theta(z(s)), \qquad \hat{m}_k = -\hat{\psi}_k, \tag{1}$$

where $z(s)$ is a latent code shared with the actor and $\rho > 0$ bounds the modulation. The modulation lets the representation reshuffle margin *between* predicates, but it cannot create safety out of violation, since $\tanh \in [-1, 1]$.

Soft conjunctions. For a budget ω on the simplex, the weighted conorm is

$$C_\omega(m) = -\tfrac{1}{\beta} \log \textstyle\sum_k \omega_k e^{-\beta m_k},$$

with the unweighted LSE as $\omega \equiv 1/K$ and the Gödel, product and Łukasiewicz conorms as the remaining standard choices (Badreddine et al., 2022). A *certificate* is a map $\mathsf{cert} : s \mapsto \{0,1\}$; we say it is γ-sound if $\mathsf{cert}(s) = 1$ implies $\max_k \psi_k \le -\gamma$, and define its *activation rate* (CAR) as the fraction of visited states with $\mathsf{cert}(s) = 1$. The two quantities trade off, and the pair (CAR, certified-but-unsafe steps) is the object we audit throughout.

The Lagrangian skeleton. All methods in this paper share one learner: PPO (Schulman et al., 2017) with two critics V_r, V_c and a dual variable λ updated by ascent on the constraint cost, which we evaluate in *certificate* form rather than as a raw penalty:

$$c(s,a) = \mathrm{relu}\left(\tau + \rho + \delta - C_\omega(m^\theta)\right), \tag{2}$$

so that $c = 0$ exactly when the (robust) certificate is active. Methods differ only in the operator family, the budget ω, the threshold τ, and whether an attribution gate enters. This isolates the contribution of each design choice.

4 METHOD

4.1 TIGHT THRESHOLDS FOR WEIGHTED SOFT CONJUNCTIONS

Let $\mathcal{A} = \mathcal{A}(\omega) = \{k : \omega_k > 0\}$ be the support of the budget and $k^\star = \arg\min_{k\in\mathcal{A}} m_k$ its binding predicate.

Lemma 1 (Weighted sandwich). *For any $m \in \mathbb{R}^K$, any $\beta > 0$ and any budget ω,*

$$\min_{k\in\mathcal{A}} m_k \;\le\; C_\omega(m) \;\le\; \min_{k\in\mathcal{A}} m_k + \frac{\log(1/\omega_{k^\star})}{\beta}. \tag{3}$$

The upper end is the best possible bound of this form: it is attained in the limit where every non-binding predicate is strictly satisfied.

Proof. Since $m_k \ge m_{k^\star}$ for all $k \in \mathcal{A}$, $\omega_{k^\star} e^{-\beta m_{k^\star}} \le \sum_{k\in\mathcal{A}} \omega_k e^{-\beta m_k} \le e^{-\beta m_{k^\star}} \sum_{k\in\mathcal{A}} \omega_k = e^{-\beta m_{k^\star}}$. Taking $-\frac{1}{\beta}\log$ and using $\sum_{k\in\mathcal{A}} \omega_k = 1$ gives both inequalities. For tightness set $m_k = m_{k^\star} + M$ for $k \ne k^\star$ and let $M \to \infty$; then $C_\omega \to m_{k^\star} + \log(1/\omega_{k^\star})/\beta$. □

The sandwich is classical for the uniform case (**?**); the weighted form shows that $\log K/\beta$ is not the correct threshold unless the budget is uniform. Suppose we wish to certify $\min_{k\in\mathcal{A}} m_k \ge \gamma$ from an observation of $C_\omega(m)$. Lemma 1 says a threshold τ works iff $\tau \ge \log(1/\omega_{k^\star})/\beta$, and that no smaller value is admissible for that budget and that state:

$$\tau_\omega \;=\; \frac{\log(1/\omega_{k^\star})}{\beta}. \tag{4}$$

Corollary 1 (Minimality). *Fix ω and β. If $\tau < \log(1/\omega_{k^\star})/\beta$ then there exist margins m with $C_\omega(m) \ge \tau + \gamma$ and $\min_{k\in\mathcal{A}} m_k < \gamma$. Hence τ_ω is the unique minimal admissible threshold.*

Proof. Take $m_{k^\star} = \gamma - \epsilon$ and $m_k = m_{k^\star} + M$ for $k \ne k^\star$, $M \to \infty$. Then $C_\omega \to \gamma - \epsilon + \log(1/\omega_{k^\star})/\beta > \tau + \gamma$ for $\epsilon < \log(1/\omega_{k^\star})/\beta - \tau$, while $\min_k m_k = \gamma - \epsilon < \gamma$. □

Theorem 1 (Soundness). *For the weighted conorm with the threshold* (4) *and any $\delta > 0$, $\rho > 0$, define*

$$\mathsf{cert}(s) = \mathbf{1}\big\{C_\omega(m^\theta(s)) \ge \tau_\omega + \rho + \delta\big\}.$$

Then for every budget, every $\beta > 0$ and every state s,

$$\mathsf{cert}(s) = 1 \implies \max_{k\in\mathcal{A}} \psi_k(s) \le -\delta.$$

Proof. By Lemma 1, $C_\omega(m^\theta) \le m^\theta_{k^\star} + \tau_\omega$, hence $\mathsf{cert} = 1$ implies $m^\theta_{k^\star} \ge \rho + \delta$. Because the sandwich upper bound holds for *every* index in the support, not only the argmin, $\mathsf{cert} = 1$ implies $m^\theta_k \ge \rho + \delta$ for all $k \in \mathcal{A}$; (1) then gives, for each such k, $m^\theta_k = -\hat{\psi}_k + \rho \tanh \varepsilon^\theta_k \le -\hat{\psi}_k + \rho$, so $\hat{\psi}_k \le -\delta$. Taking the worst index gives the claim. □

Theorem 1 holds for the uniform budget too, with $\tau = \log K/\beta$; the point is that this is strictly larger whenever $\omega_{k^\star} > 1/K$, which is the generic case for a weighted budget. The difference is not a constant factor: with β adapted so that the conorm's gradient remains informative, $\log K/\beta$ is comparable to the whole margin spread of the population, and the certificate becomes vacuous (Section 5). Formally, fixing the support and the budget map,

$$\mathsf{cert}^{\mathrm{unif}}(s) = 1 \implies \mathsf{cert}(s) = 1 \quad \text{whenever } \omega_{k^\star} \ge 1/K, \tag{5}$$

and the strictness is $\log(K\omega_{k^\star})/\beta$ of unnecessary robustness.

4.2 BINDINGNESS IS NOT CONTROL AUTHORITY

The bound (4) rewards concentrating budget on the binding predicate: with $\omega_{k^\star} \to 1$ the certificate loses no slack at all, and C_ω becomes the exact minimum. A budget must however be a function of the state that is computable online. The conorm supplies one for free, its own weight $w_k = \mathrm{softmax}_k(-\beta m_k)$, and this is what temperature-controlled LSE conjunctions adapt (**?**). It has a systematic failure mode. *Bindingness is associational*: a predicate is weighted by how small its margin is, not by whether any action could have made it larger. An agent-independent monitor — one whose residual is a function of exogenous state only — is therefore weighted *maximally* precisely when it is violated, i.e. when the agent can do nothing about it.

We instead attribute by intervention. Write $J_k(s,a) = \partial\psi_k(s,a)/\partial a \in \mathbb{R}^d$ for the action sensitivity of the residual, and order predicates by the causal order of the conjunction, $j \prec k$ iff $m_j < m_k$ (ties broken arbitrarily but consistently).

Definition 1 (IARS attribution)**.** With σ the logistic function, $\varsigma > 0$ the ramp width and μ the authority threshold, define

$$\mathrm{ctrl}_k = \sigma\Big(\frac{\|J_k\|/\operatorname{med}_j \|J_j\| - \mu}{\varsigma}\Big), \qquad \text{(authority)} \tag{6}$$

$$\mathrm{indep}_k = \mathrm{relu}\Big(\frac{c_1 - \max_{j\prec k} \cos(J_k, J_j)}{c_1 - c_0}\Big) \wedge 1, \qquad \text{(non-redundancy)} \tag{7}$$

$$\mathrm{cov}_k = \mathrm{ctrl}_k \cdot \mathrm{indep}_k, \tag{8}$$

with $\cos(u,v) = \langle u, v\rangle/(\|u\|\|v\|)$ and $0 < c_0 < c_1 < 1$. The gate $\mathrm{cov}_k \in [0,1]$ is the product of the predicate's relative action authority, measured against the median authority in the specification, and its non-redundancy with respect to every more binding predicate. Appendix F lists the constants used throughout.

Lemma 2 (Invariances)**.** (i) Exogenous monitors. *If $\partial\psi_k/\partial a \equiv 0$ on a neighbourhood of the data then $\widehat{\mathrm{ctrl}}_k \to 0$ as $\mu/\varsigma \to \infty$ and hence $\mathrm{cov}_k \to 0$.* (ii) Redundant restatements. *If $\psi_k = \psi_j - c$ with $c > 0$ constant, equivalently $m_k = m_j + c$, so that j is strictly more binding at every state, and $J_k \parallel J_j$ with positive orientation, then $\cos(J_k, J_j) = 1$, hence $\mathrm{indep}_k = 0$ and $\mathrm{cov}_k = 0$; predicate k is also* implied *by j, since $\psi_j \le -\gamma$ gives $\psi_k = \psi_j - c \le -\gamma - c \le -\gamma$.*

Proof. (i) $\|J_k\| = 0$ makes the ratio 0 and the logistic argument $-\mu/\varsigma$, which vanishes as $\mu/\varsigma \to \infty$. (ii) By $m_k = m_j + c > m_j$ the ordering $j \prec k$ is state-independent, and collinearity with positive orientation makes the maximum in Definition 1 equal to 1, so the clip saturates at 0. The implication is immediate from $\psi_k = \psi_j - c$. □

Why condition on the causal order? The natural residualisation would project J_k against the span of all other $\{J_j\}_{j\neq k}$ and use the norm of the residual. That definition is order-free, but in a d-dimensional action space at most d predicates can retain non-zero residual, so for $K \gg d$ it degenerates to a rank-d selection and stops discriminating between predicates that are genuinely independent but lie in a shared low-dimensional image — which is the common case for geometric predicates. Conditioning on the *more binding* predicates instead asks the question the conjunction actually asks: does k carry authority that the predicates already dominating the conorm do not? This keeps the rank constraint from biting, gives a state-independent answer for affinely related predicates (Lemma 2(ii)), and yields the gate we need. Appendix D reports the degenerate behaviour of the order-free variant.

4.3 A SOUND CERTIFICATE FOR A SPLIT SPECIFICATION

Let the specification be partitioned by the gate at a threshold $\zeta \in (0,1)$ into

$$\mathcal{A} = \{k : \mathrm{cov}_k > \zeta\}, \qquad \mathcal{M} = \{k : \mathrm{ctrl}_k < \zeta\}, \qquad \mathcal{D} = \{k : \mathrm{indep}_k < \zeta,\ \mathrm{ctrl}_k \ge \zeta\}, \tag{9}$$

the *attributable*, *monitored* and *dropped* predicates. A budget is supported on $\mathcal{A}$,

$$\omega_k(s) \ \propto\ w_k(s)\,\mathrm{cov}_k \quad (k \in \mathcal{A}), \qquad \sum_{k\in\mathcal{A}} \omega_k = 1, \tag{10}$$

and the deployment certificate is the conjunction of a soft, robustness-margined statement about $\mathcal{A}$ and an exact, margin-free statement about $\mathcal{M}$:

$$\mathsf{cert}^{\mathrm{CT}}(s) = \mathbf{1}\Big\{ \underbrace{C_\omega(m^\theta(s)) \;\geq\; \tau_\omega + \rho + \delta}_{\text{attributable part}} \Big\} \;\wedge\; \mathbf{1}\Big\{ \underbrace{\max_{k\in\mathcal{M}} \hat{\psi}_k(s) \leq -\delta_m}_{\text{exact monitor}} \Big\}. \tag{11}$$

Theorem 2 (Soundness of the decomposition)**.** *Assume the residuals are exact ($\hat{\psi} = \psi$) and the gate classifies each predicate correctly, in the sense that $\mathcal{A}$ contains no exogenous predicate, $\mathcal{M}$ contains every predicate with $\|J_k\| \equiv 0$, and every $k \in \mathcal{D}$ admits a $j \prec k$ with $j \in \mathcal{A} \cup \mathcal{M}$ and $\psi_k = \psi_j - c_k$ for some constant $c_k \geq 0$. Then for any budget supported on $\mathcal{A}$, $\mathsf{cert}^{\mathrm{CT}}(s) = 1 \Rightarrow \max_{k\leq K} \psi_k(s) \leq -\min(\delta, \delta_m) \leq 0$: the deployment certificate is $\min(\delta, \delta_m)$-sound for the* full *specification, not only for $\mathcal{A} \cup \mathcal{M}$.*

Proof. For $k \in \mathcal{A}$, Theorem 1 gives $\psi_k \leq -\delta$. For $k \in \mathcal{M}$, the exact monitor gives $\psi_k \leq -\delta_m$. For $k \in \mathcal{D}$, follow the dominator chain: pick $j \prec k$ as in the hypothesis and repeat while the dominator is itself in $\mathcal{D}$. The order $\prec$ is a strict order (strictly smaller margins) and hence acyclic on a finite set, so the chain terminates at some $\ell \in \mathcal{A} \cup \mathcal{M}$ with $\psi_k = \psi_\ell - \sum_i c_i \leq \psi_\ell \leq -\min(\delta, \delta_m)$. Taking the worst index over $k \leq K$ gives the claim. □

Proposition 1 (Activation rate)**.** *Fix the support $\mathcal{A}$ and a budget map with $\inf_s \omega_{k^\star(s)}(s) \geq b > 0$. Write $\overline{\mathrm{CAR}}_\infty = \mathbb{P}_s[\max_{k\in\mathcal{A}} \psi_k \leq -\delta]$ for the activation rate of the exact (non-differentiable) conjunction over $\mathcal{A}$ with the modulation margin ρ stripped out. Then for every $\beta > 0$ and every budget,*

$$\mathbb{P}_s\Big[\max_{k\in\mathcal{A}} \psi_k \leq -2\rho - \delta - \tau_\omega(s) \Big] \;\leq\; \mathrm{CAR}(\beta) \;\leq\; \overline{\mathrm{CAR}}_\infty. \tag{12}$$

The two endpoints differ by at most ρ, the worst-case modulation budget: they coincide when the modulation saturates ($\tanh\varepsilon \equiv +1$ on $\mathcal{A}$) and jointly converge as $\tau_\omega \to 0$. Consequently the gap between the achievable and the unweighted thresholds, τ_ω versus $\log K/\beta$, is a lower bound on the activation rate lost by the uniform certificate.

Proof. Upper bound. By Lemma 1, $\mathsf{cert} = 1$ implies $\min_{k\in\mathcal{A}} m_k^\theta \geq \rho + \delta$. Hence *every* $k \in \mathcal{A}$ satisfies $m_k^\theta \geq \rho + \delta$, and since $m_k^\theta = -\hat{\psi}_k + \rho \tanh \varepsilon_k^\theta \leq -\hat{\psi}_k + \rho$, every $k \in \mathcal{A}$ satisfies $\hat{\psi}_k \leq -\delta$. Taking the worst index gives the event on the right.

Lower bound. The same sandwich is also a *sufficient* condition. If $\max_{k\in\mathcal{A}} \hat{\psi}_k \leq -2\rho - \delta - \tau_\omega(s)$ then, for every $k \in \mathcal{A}$, $m_k^\theta \geq -\hat{\psi}_k - \rho \geq \rho + \delta + \tau_\omega(s)$, so $\min_{k\in\mathcal{A}} m_k^\theta \geq \rho + \delta + \tau_\omega$ and therefore $C_\omega \geq \min_{k\in\mathcal{A}} m_k^\theta \geq \tau_\omega + \rho + \delta$, i.e. $\mathsf{cert} = 1$. The extra ρ here is the worst-case under-estimate of $-\hat{\psi}$ by the modulation, which is exactly what separates the two endpoints; it does not shrink with β, whereas the $O(1/\beta)$ slack of the conorm does, so along any sequence with $\tau_\omega \to 0$ both bounds collide with the corresponding exact-conjunction rates. □

Proposition 1 is the formal statement of the paper's design principle: the activation rate of a *sound* certificate is controlled by the threshold, and the threshold is controlled by two design choices — where the budget sits, and which margin the non-attributable constraints are held to. An exact nominal check contributes $\delta_m = 0$ to the required margin instead of $\rho + \delta$, and the difference is a strictly larger set of certified states at unchanged soundness. Note also that (2)—(3) require no gradient through the monitor: there is no reason to relax a constraint no action can move.

4.4 Algorithm

Algorithm 1 (Appendix A) instantiates the construction on the PPO-Lagrangian skeleton. The gate is estimated every F iterations from a minibatch of rollouts by automatic differentiation through the environment's one-step residual map, then smoothed by an EMA; the budget is recomputed per state from the gate and the conorm's own weights; the cost channel is the certificate residual (2). Total overhead is one extra backward pass per F iterations through a $K \times d$ Jacobian, which is negligible relative to the PPO update (measured in Appendix F).

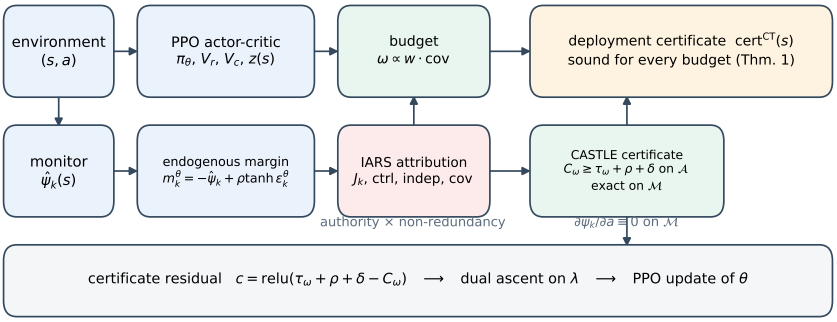


Figure 1: CASTLE inside a PPO-Lagrangian learner. Predicate residuals $\hat{\psi}$ feed both the endogenous margin (1) and the IARS attribution, which differentiates the one-step residual map to obtain per-predicate action Jacobians J_k. Authority ctrl and non-redundancy indep define the gate cov, which splits the specification into an attributable set (soft conorm with the minimal threshold τ_ω), a monitored set (exact nominal check) and a dropped set of implied restatements. Both certificate channels feed the dual variable through the residual (2).

5 Experiments

Environments. All three environments are vectorised pure-PyTorch simulators with analytic one-step residual maps, so that the action Jacobian J_k is available by autodifferentiation without a model. *SafeNavC* is a planar double integrator with drag, three bouncing obstacles and four wall predicates ($K{=}8$), with one obstacle parked 0.30–0.50 from the goal so that reaching it forces a clearance/performance trade-off. *SafeReachC* is a two-link planar manipulator with six predicates (joint limits, rate limits, obstacle clearance and end-effector speed). *SafeConf* reuses SafeNavC's physics and degrades the *specification*: predicates 3 and 4 restate predicate 2, so their Jacobians are co-oriented and they add no control authority, and predicate 8 is the clearance between a bouncing co-worker and a fixed workstation, a function of the obstacle trajectory alone, hence $\partial\psi_8/\partial a \equiv 0$. Ground truth $\mathrm{gt}_k \in \{0, 1\}$ marks the five predicates with independent authority. The offsets exceed 2ρ, so the domination is state-independent and attribution is well posed at every state.

Protocol. Eight methods — PPO-Lagrangian, reward shaping, safety layer, Gödel, product and Łukasiewicz t-norms, DENSSE-style uniform LSE, and CASTLE — and five CASTLE ablations, each on three environments (ablations on two) with three seeds, $200{,}000$ environment steps, and one shared hyper-parameter set (Appendix F). Reported quantities are success rate (SR), cumulative agent-controllable cost, violation rate (VR), and, from a certificate audit over 32 fresh episodes, the activation rate (CAR), certified-but-unsafe steps (CBU, a certificate fires while a controllable residual is positive), CBU over *all* predicates ($\mathrm{CBU}_{\mathrm{all}}$), and the attribution AUROC against gt.

5.1 Main results

Table 1 reports the main grid. Three observations drive the rest of the paper. First, on *SafeConf* the uniform-threshold LSE certificate of prior work fires on 0.215 of states, while CASTLE's fires on 0.381 — a factor of 2 — and neither ever certifies an unsafe state; the gap is the $\log(K\omega_{k^\star})/\beta$ slack of (5) plus the margin posted against the agent-independent monitor. Second, the product and Łukasiewicz t-norms, at their natural threshold $\tau = 0$, certify 1463 steps at which a controllable residual is positive: their smoothed conjunctions can exceed $\rho + \delta$ while one argument is negative, and no threshold repairs this. Third, CASTLE is never worse on the safety–performance pair and is best on *SafeConf* (0.885 success at the lowest controllable cost), which is the environment where attribution carries information. Learning curves and the safety–performance frontier are shown in Appendix A.

Table 1: Main results (mean over three seeds; ± standard error). SR: success rate; Cost: cumulative agent-controllable cost per episode; VR: violation rate; CAR: certificate activation rate; CBU: certified-but-unsafe steps summed over the audit; $\mathrm{CBU}_{\mathrm{all}}$: same but counting *every* predicate, including agent-independent ones. Bold marks the best value per environment. Every method except the product and Łukasiewicz t-norms is sound on the controllable predicates ($\mathrm{CBU} = 0$).

Env	Method	SR ↑	Cost↓	VR ↓	CAR ↑	CBU	$\mathrm{CBU}_{\mathrm{all}}$
SafeNavC	PPO-Lag	0.93±0.06	0.29±0.06	0.34±0.05	0.840	0	0
	RewShap	0.94±0.03	0.26±0.04	0.35±0.04	0.857	0	0
	SafetyLayer	0.53±0.06	0.46±0.08	0.64±0.05	0.699	0	0
	Gödel	0.86±0.14	0.33±0.06	0.40±0.05	0.844	0	0
	Product	0.96±0.02	**0.24±0.01**	0.31±0.04	1.000	518	259
	Łukasiewicz	0.75±0.04	0.43±0.02	0.40±0.01	0.000	0	0
	DENSSE-LSE	0.77±0.11	0.44±0.10	0.49±0.05	0.858	0	0
	CASTLE	**0.97±0.02**	0.27±0.03	0.33±0.06	0.858	0	0
SafeReachC	PPO-Lag	0.50±0.04	2434.03±359.24	0.98±0.01	0.210	0	0
	RewShap	0.11±0.02	**97.38±33.52**	0.81±0.13	0.700	0	0
	SafetyLayer	0.22±0.07	212.28±36.04	0.97±0.02	0.245	0	0
	Gödel	0.38±0.04	2616.33±224.13	0.96±0.01	0.200	0	0
	Product	0.42±0.04	2138.63±455.46	0.97±0.00	0.266	562	281
	Łukasiewicz	**0.52±0.07**	2440.95±419.72	0.97±0.02	0.000	0	0
	DENSSE-LSE	0.49±0.07	1952.95±362.66	0.95±0.02	0.200	0	0
	CASTLE	0.47±0.07	2339.51±392.26	0.98±0.01	0.227	0	0
SafeConf	PPO-Lag	0.97±0.02	**0.09±0.04**	0.20±0.04	0.271	0	0
	RewShap	0.91±0.07	0.12±0.04	0.23±0.05	0.268	0	0
	SafetyLayer	0.94±0.05	0.13±0.03	0.21±0.02	0.284	0	0
	Gödel	0.98±0.02	0.14±0.02	0.22±0.04	0.281	0	0
	Product	**0.99±0.01**	0.10±0.03	0.20±0.02	0.658	975	923
	Łukasiewicz	0.89±0.08	0.15±0.06	0.20±0.06	0.000	0	0
	DENSSE-LSE	0.97±0.02	0.11±0.01	0.21±0.05	0.215	0	0
	CASTLE	0.89±0.11	0.12±0.03	0.20±0.05	0.381	0	0

5.2 Is the certificate sound, and is it informative?

A certificate that never fires is sound and useless; one that always fires is informative and unsound. Figure 5a (Appendix A) sweeps β on trained policies and reports CAR together with the two endpoints of Proposition 1: the sufficient-condition lower bound (dotted) and the exact-conjunction ceiling $\overline{\mathrm{CAR}}_\infty$ (dashed). CASTLE's CAR sits inside the predicted band at every β and coincides with the ceiling $\overline{\mathrm{CAR}}_\infty$ once $\tau_\omega < \delta$, while the uniform-threshold certificate stays far below the band's lower edge at every β — a gap that raising β alone cannot repair, since $\log K/\beta$ shrinks at the same rate but from a K-times larger constant, the empirical content of Corollary 1. The realised modulation $\mathbb{E}[\rho \tanh \varepsilon^\theta]$ on the support is 76 of the worst case ρ: the learner drives the modulation towards its saturating end rather than paying the worst case.

Figure 5b is the ablation ladder at fixed β, ordered by the mechanism removed (complete per-variant numbers in Appendix A). Dropping the exact monitor raises CAR to 0.950, past the sound ceiling, and produces 3253 certified-but-unsafe steps. Dropping the causal gate — the uniform conorm over all eight predicates with the same tight threshold — is sound but loses 25% of its activation rate, because the margin $\rho + \delta$ must then be posted against the agent-independent monitor. Replacing τ_ω by the legacy $\log K/\beta$ loses a further 0.042. Removing the endogenous modulation is comparable on activation but degrades the safety–performance pair, since the modulation is the only channel through which the representation can reshuffle margin between predicates before the actor has learned to do so. The ladder is monotone in exactly the direction predicted by Theorem 1 and Proposition 1 — the strongest empirical support for the theory.

Two channels, and an infeasible specification. The composite certificate fires only when both channels agree, so its activation rate is the *minimum* of the two. On *SafeConf* the soft channel over the five attributable predicates fires on 0.950 of them, whereas the exact monitor passes on fewer; the composite rate is 0.381, and the difference measures the monitor, not the threshold. Sampling

2048 resets shows that on *SafeConf* a strictly safe *and* successful state exists on only 9.9% of resets, against 100.0% on *SafeNavC*, and 100.0% when restricted to the predicates the action can actually move. The infeasibility is therefore entirely exogenous, and a sound certificate is *right* to stay below 1 here: CASTLE certifies the states it can move and reports the rest rather than averaging them into a margin. The uniform-threshold certificate conflates the two channels — it must post $\log K/\beta$ against the monitor on every state — which is why it lands at 0.215.

5.3 Does the attribution identify control authority?

Table 3 and Figure 6 (Appendix A) evaluate attribution directly against the ground truth of *SafeConf*. Ranked by the conorm's own softmax weights — the associational attribution that the temperature controller implicitly optimises — the agent-independent monitor is the *highest*-scoring predicate, so AUROC falls to 0.53 and, at the best threshold, the confusion matrix has 3 false positives. The gate inverts this: AUROC 1.00 at the predicate level and 0.95 over states. The ablation *ctrl-only* keeps the two restatements in the support, spending 8% of the budget on them — exactly the slack that the full gate removes. The redundancy channel is not decorative.

5.4 When does attribution matter?

On *SafeNavC* and *SafeReachC* every predicate has independent authority, the gate converges to one, and CASTLE reduces to DENSSE-style LSE with a slightly tighter threshold; the differences there are within seed noise (Table 1). This delimits the contribution: attribution is worth its overhead exactly when the specification contains constraints the agent cannot move. Our two failure modes — an exogenous monitor and a redundant restatement — are common in embodied specifications (a human co-worker, a thermal or regulatory margin, a limit restated in different units), and both are invisible to margin-based weighting.

6 Discussion

Two caveats bound what we claim. First, the decomposition of (11) is only as sound as the gate: Theorem 2 assumes the attribution classifies predicates correctly, and a misclassified exogenous predicate moved into $\mathcal{A}$ would silently cost soundness. In our runs the gate separates cleanly with an order-of-magnitude gap (Figure 2, Appendix A), but a monitor-facing deployment should audit the gate itself, e.g. by checking that ctrl_k stays low over a rollout window before dropping a predicate from $\mathcal{A}$. Second, we evaluate in simulation with analytic residual maps. Where residuals come from vision or a learned monitor, the derivative must be estimated and its error propagates into the gate and threshold; a conservative default that only demotes predicates whose estimated authority is below ζ on a whole rollout would degrade gracefully. We leave derivative error to future work.

The broader point is that the smooth conjunction is doing two jobs, and that the second one is under-designed. As a penalty it wants to be smooth and to spread gradient over the predicates that can still be improved. As a certificate it wants to approach the exact minimum while remaining sound, and to hold the right margin on the right constraints. The threshold is determined by the budget in closed form, and the budget should be decided by control authority rather than by which margin happens to be small.

Reproducibility statement

All results are produced by a single entry point (`run_suite.py`) from fixed seeds; the environments, methods, ablations, certificate audit and figure generation are self-contained PyTorch, requiring neither MuJoCo nor a GPU. The theoretical results are stated with complete proofs in Appendix B. The feasibility diagnostic of Section 5.2 and the attribution-cost measurement of Appendix F are stand-alone scripts that write the numbers they report to disk. Source code, configuration and the exact seeds for every run are released with the supplementary material.

AI USE STATEMENT

In this work, generative AI tools were used for implementing the experimental code, executing the numerical experiments, generating the figures, and drafting and editing parts of the manuscript. The research direction, the theoretical framework and the decision to submit were made by the authors. All mathematical statements were checked against their proofs, all reported numbers were regenerated from the released code at fixed seeds, and every reference was verified against its primary source before inclusion. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Andrea Agazzi, Vittorio Carlei, Marco Romito, and Samuele Saviozzi. Global optimization via softmin energy minimization. *arXiv:2509.17815*, 2025.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *European Control Conference (ECC)*, 2019.

Greg Anderson, Abhinav Verma, Isil Dillig, and Swarat Chaudhuri. Neurosymbolic reinforcement learning with formally verified exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Anonymous. Differentiable endogenous neuro-symbolic safe exploration: a log-sum-exp barrier for PPO-Lagrangian constrained reinforcement learning. *Under review*, 2027.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022.

Hongye Cao, Fan Feng, Tianpei Yang, Jing Huo, and Yang Gao. Causal information prioritization for efficient reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Chenyang Cao, Yucheng Xin, Silang Wu, Longxiang He, Zichen Yan, Junbo Tan, and Xueqian Wang. FOSP: Fine-tuning offline safe policy through world models. In *International Conference on Learning Representations (ICLR)*, 2025.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2017.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv:1801.08757*, 2018.

Quentin Delfosse, Hikaru Shindo, Devendra Dhami, and Kristian Kersting. Interpretable and explainable logical policies via neurally guided symbolic abstraction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In *AAAI Conference on Artificial Intelligence*, 2018.

Qiao Gu, Yuanliang Ju, Shengxiang Sun, Igor Gilitschenski, Haruki Nishimura, Masha Itkina, and Florian Shkurti. SAFE: Multitask failure detection for vision-language-action models. *arXiv:2506.09937*, 2025.

Edwin Hamel-De le Court, Francesco Belardinelli, and Alexander W. Goodall. Probabilistic shielding for safe reinforcement learning. *arXiv:2503.07671*, 2025.

Songqiao Hu, Zeyi Liu, Shuang Liu, Jun Cen, Zihan Meng, Shihefeng Wang, Xiang Li, and Xiao He. VLSA: Vision-language-action models with plug-and-play safety constraint layer. *arXiv:2512.11891*, 2025.

Yuzheng Hu, Fan Wu, Haotian Ye, David Forsyth, James Zou, Nan Jiang, Jiaqi W. Ma, and Han Zhao. A snapshot of influence: A local data attribution framework for online reinforcement learning. *arXiv:2505.19281*, 2025.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Juntao Dai, and Yaodong Yang. Safety-Gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Ankita Kushwaha, Kiran Ravish, Preeti Lamba, and Pawan Kumar. A survey of safe reinforcement learning and constrained MDPs: A technical survey on single-agent and multi-agent safety. *arXiv:2505.17342*, 2025.

Jianming Ma, Rongjun Jin, Xiaxi Si, Yang Zhang, Yiheng Li, and Yue Gao. ActSafeGuard: Differentiable and training-aligned constraint enforcement for flow-matching policies. *arXiv:2609.11697*, 2026.

Thomas Mesnard, Théophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, and Rémi Munos. Counterfactual credit assignment in model-free reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2021.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Victor Pfanschilling, Hikaru Shindo, Devendra Singh Dhami, and Kristian Kersting. NeST: The neuro-symbolic transpiler. *International Journal of Approximate Reasoning*, 179:109369, 2025.

Pedram Rabiee and Amirsaeid Safari. Safe exploration in reinforcement learning: Training backup control barrier functions with zero training-time safety violations. In *Learning for Dynamics and Control Conference (L4DC), Proceedings of Machine Learning Research*, 2025.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *OpenAI Technical Report*, 2019.

Amirsaeid Safari and Jesse B. Hoagg. Safe navigation in unmapped environments for robotic systems with input constraints. *arXiv:2410.02106*, 2025.

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

Hikaru Shindo, Viktor Pfanschilling, Devendra Singh Dhami, and Kristian Kersting. Learning differentiable logic programs for abstract visual reasoning. *Machine Learning*, 113(11):8533–8584, 2024.

Hikaru Shindo, Quentin Delfosse, Devendra Singh Dhami, and Kristian Kersting. BlendRL: Merging symbolic and neural policy learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Arseny Skryagin, Daniel Ochs, Phillip Deibert, Simon Kohaut, Devendra Singh Dhami, and Kristian Kersting. Answer set networks: Casting answer set programming into deep learning. *arXiv:2412.14814*, 2024.

Kai Tang, Peidong Jia, Zhong Chu, Jixian Wu, Rui Ma, Jiajun Cao, Fangyuan Zhao, Sixiang Chen, Yichen Guo, Xiaowei Chi, Chun-Kai Fan, Kevin Zhang, Jinchang Xu, Fubing Yang, Weishi Mi, Xiaozhu Ju, Jian Tang, and Shanghang Zhang. SafeDojo: Safe reinforcement learning for VLA via interactive world model. *arXiv:2606.20698*, 2026.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations (ICLR)*, 2019.

Chenggang Wang, Xinyi Wang, Yutong Dong, Lei Song, and Xinping Guan. Multi-constraint safe reinforcement learning via closed-form solution for log-sum-exp approximation of control barrier functions. *arXiv:2505.00671*, 2025.

Lizhi Yang, Blake Werner, Ryan K. Cosner, David Fridovich-Keil, Preston Culbertson, and Aaron D. Ames. SHIELD: Safety on humanoids via CBFs in expectation on learned dynamics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

A Additional experimental results

B Proofs omitted from the main text

Lemma 1. Both inequalities follow from $\omega_{k^\star} e^{-\beta m_{k^\star}} \le \sum_{k\in\mathcal{A}} \omega_k e^{-\beta m_k} \le e^{-\beta m_{k^\star}} \sum_{k\in\mathcal{A}} \omega_k$. The first inequality uses $m_k \ge m_{k^\star}$ and $\omega_k \le 1$; the second uses $e^{-\beta m_k} \le e^{-\beta m_{k^\star}}$ for every $k \in \mathcal{A}$ and $\sum \omega = 1$. The limiting construction shows the upper bound cannot be improved by a constant independent of m: for $m_k = m_{k^\star} + M$ the ratio $\sum_k \omega_k e^{-\beta(m_k - m_{k^\star})} \to \omega_{k^\star}$ as $M \to \infty$, so $C_\omega \to m_{k^\star} + \log(1/\omega_{k^\star})/\beta$.

Algorithm 1 CASTLE on a PPO-Lagrangian skeleton

Require: environments, predicate residuals $\hat{\psi}$, gate params $(\mu, \varsigma, c_0, c_1, \zeta)$, margins ρ, δ, δ_m, attribution period F, budget mix ϵ

init actor-critic, dual $\lambda \leftarrow 0$, temperature $\beta \leftarrow \beta_0$, gates $\mathrm{cov}, \mathrm{ctrl}, \mathrm{indep} \leftarrow 1$
for iteration $= 1, 2, \ldots$ **do**
  collect N rollouts with π_θ; store $s, a, \hat{\psi}(s)$
  if iteration $\bmod F = 0$ **then**
    $J_k \leftarrow \partial \hat{\psi}_k(s')/\partial a$ by autodiff through the one-step map (batch)
    $(\mathrm{cov}', \mathrm{indep}', \mathrm{ctrl}') \leftarrow \mathrm{IARS}(J_k, m^\theta)$ ▷ Definition 1
    update EMA gates; set $\mathcal{A}, \mathcal{M}, \mathcal{D}$ from $\mathrm{cov}, \mathrm{ctrl}, \mathrm{indep}$ and ζ
  end if
  $k^\star(s) \leftarrow \arg\min_{k \in \mathcal{A}} m_k^\theta$; $\omega_k(s) \leftarrow (1-\epsilon)\, \frac{w_k \mathrm{cov}_k}{\sum_{j \in \mathcal{A}} w_j \mathrm{cov}_j} + \epsilon/|\mathcal{A}|$ on $\mathcal{A}$ ▷ Eq. (10)
  $c(s,a) \leftarrow \mathrm{relu}\left(\tau_\omega(s) + \rho + \delta - C_\omega(m^\theta(s))\right)$ ▷ Eqs. (4), (2)
  update β by entropy feedback on smin weights; update $\lambda \leftarrow \lambda + \eta_\lambda \bar{c}$ (clip)
  PPO update of θ with advantage $\hat{A}_r$ and penalty $\lambda\, c$
  deploy: $\mathrm{cert}^{\mathrm{CT}}$ as in (11); monitor $\mathcal{M}$ exactly
end for

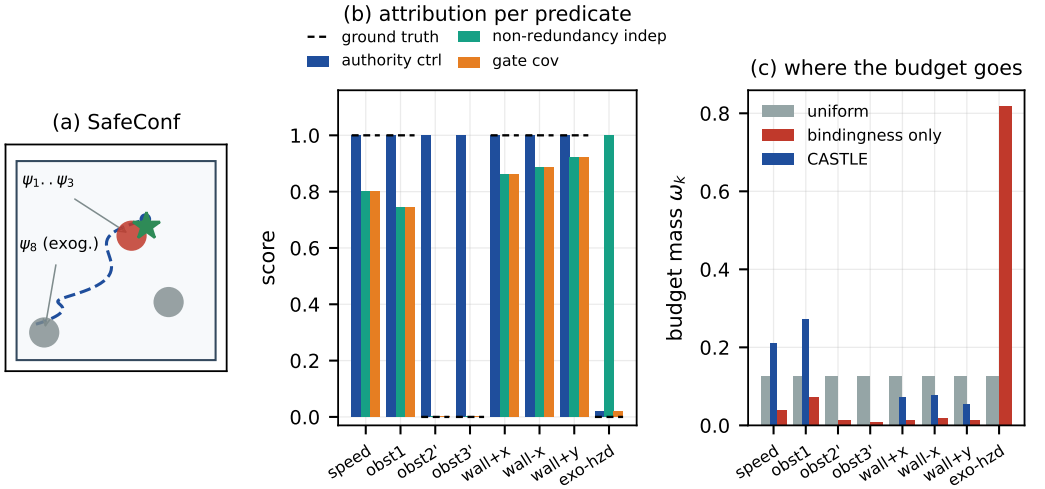


Figure 2: Left: *SafeConf*. Predicates $3, 4$ restate predicate 2; predicate 8 is an agent-independent co-worker clearance. Middle: measured authority, non-redundancy and gate per predicate (256 states). The exogenous monitor has $\mathrm{ctrl} \approx 0.02$ and the two restatements have $\mathrm{indep} = 0$, while independent predicates score 0.84; the resulting AUROC against ground truth is 1.00 for the gate against 0.53 for the conorm's own softmax weights. Right: where the budget goes. The associational budget spends 82% of its mass on the agent-independent monitor and 2% on restatements; CASTLE spends 0% and 0%.

Corollary 1. As in the main text; note the construction uses arbitrarily well-satisfied non-binding predicates, which is the regime that any magnitude-blind threshold must handle.

Theorem 1. Lemma 1 gives $C_\omega \le \min_{k \in \mathcal{A}} m_k^\theta + \tau_\omega$; substituting $\mathrm{cert} = 1$ yields $\min_{k \in \mathcal{A}} m_k^\theta \ge \rho + \delta$, i.e. the lower end of the sandwich is above the threshold and therefore *all* margins in the support pass it, not only the binding one: $m_k^\theta \ge \rho + \delta$ for every $k \in \mathcal{A}$. Since $m_k^\theta = \hat{m}_k + \rho \tanh \varepsilon_k^\theta \le \hat{m}_k + \rho$ and $\hat{m}_k = -\hat{\psi}_k$, each $k \in \mathcal{A}$ satisfies $\hat{\psi}_k \le -\delta$, and the worst index gives $\max_{k \in \mathcal{A}} \hat{\psi}_k \le -\delta$. The argument does not use any property of the budget beyond $\sum_{k \in \mathcal{A}} \omega_k = 1$ and $\mathcal{A} \neq \emptyset$.

Theorem 2. Partition $[K]$ into $\mathcal{A}, \mathcal{M}, \mathcal{D}$. For $k \in \mathcal{A}$, Theorem 1 gives $\psi_k \le -\delta$. For $k \in \mathcal{M}$, $\|J_k\| \equiv 0$ and the monitor condition gives $\psi_k \le -\delta_m$. For $k \in \mathcal{D}$ there is j with $j \prec k$ and

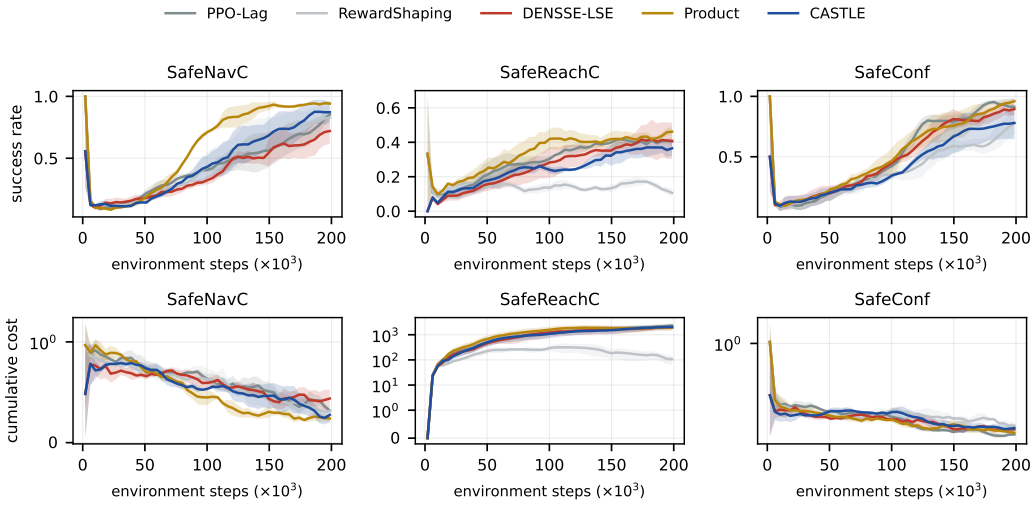


Figure 3: Learning curves (mean ± s.e. over three seeds). Top: success rate. Bottom: cumulative agent-controllable cost. CASTLE reaches higher success at lower cost on *SafeConf*; on *SafeNavC* and *SafeReachC*, where every predicate is attributable and the gate is inactive, it is statistically indistinguishable from the uniform LSE baseline, as it should be.

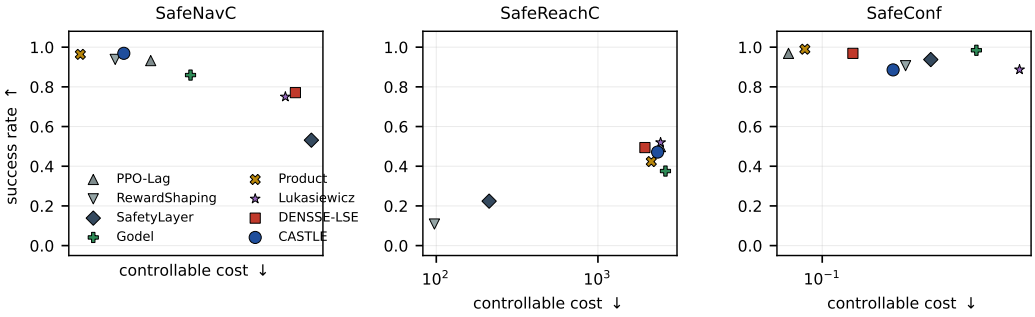


Figure 4: Safety–performance frontier (success rate against controllable cost; up and to the left is better). Each point is one (method, environment) cell averaged over seeds. The product and Łukasiewicz t-norms reach competitive success rates but at higher cost, and their certificates are the unsound ones (Table 1). The safety layer is bimodal: where the safe set is fat (*SafeConf*) its projection rarely has to fight the objective and it is competitive, but where clearance must be traded against progress (*SafeNavC*, *SafeReachC*) it gives up success; it produces no certificate in either regime.

$\psi_k = \psi_j + c_k$, $c_k \geq \rho + \delta > 0$; iterating the relation produces a strictly decreasing sequence of margins, which on a finite set must terminate in $\mathcal{A} \cup \mathcal{M}$. Along the chain ψ decreases, so the bound transfers, and the smallest guaranteed slack is $\min(\delta, \delta_m)$. Hence $\max_k \psi_k \leq -\min(\delta, \delta_m)$ and in particular the certificate implies $\max_k \psi_k \leq 0$ whenever $\delta_m \geq 0$.

Proposition 1. For the upper bound, Lemma 1 plus (1) applied to every index of the support gives $\mathsf{cert} = 1 \Rightarrow m_k^\theta \geq \rho + \delta \; \forall k \in \mathcal{A} \Rightarrow \max_{k \in \mathcal{A}} \psi_k \leq -\delta$. For the lower bound, the same computation read in the sufficient direction shows that $\max_{k \in \mathcal{A}} \hat{\psi}_k \leq -2\rho - \delta - \tau_\omega$ implies $m_k^\theta \geq -\hat{\psi}_k - \rho \geq \rho + \delta + \tau_\omega$ for every $k \in \mathcal{A}$, hence $\min_{k \in \mathcal{A}} m_k^\theta \geq \rho + \delta + \tau_\omega$. Since the conorm dominates the minimum on its support, $C_\omega \geq \min_{k \in \mathcal{A}} m_k^\theta \geq \rho + \delta + \tau_\omega$ and $\mathsf{cert} = 1$. The two thresholds differ by exactly ρ, the worst-case modulation budget; they coincide when $\tanh \varepsilon \equiv +1$ on $\mathcal{A}$. Finally, as $\beta \to \infty$ with a budget bounded below on its support, $\tau_\omega \to 0$ and $C_\omega \to \min_{k \in \mathcal{A}} m_k^\theta$, so the two

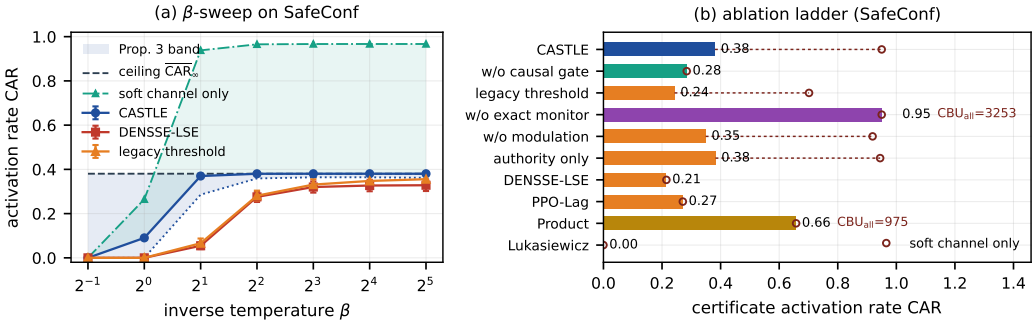


Figure 5: (a) Certificate activation rate against inverse temperature on *SafeConf*, averaged over seeds, with the two endpoints of Proposition 1: the sufficient-condition lower bound (dotted) and the ceiling $\overline{\mathrm{CAR}}_\infty$ (dashed). CASTLE's CAR lies inside the band and converges to the ceiling; the uniform-threshold certificate stays below the lower edge at every β. The dash-dotted curve is the soft channel alone, before the exact monitor is conjoined; the shaded strip between the two is the cost of the agent-independent clearance, and it is what makes the composite rate a statement the agent can honestly make. (b) Ablation ladder: CAR (bars, left axis), the soft channel (open circles) and certified-but-unsafe steps (numbers on the bars, $\mathrm{CBU}_{\mathrm{all}}$). Only the variant that removes the exact monitor becomes unsound; every sound variant respects the ordering implied by the thresholds it uses.

Table 2: Ablation ladder on *SafeConf* (mean over three seeds; ± s.e.). Rows are ordered by the mechanism removed from the full method. CAR is the composite certificate; $\mathrm{CAR}_{\mathrm{soft}}$ drops the exact monitor and is the quantity the latter variant optimises; $\mathrm{CAR}_{\mathrm{th}}$ is the predicted lower bound $\mathbb{P}[\max_{k\in\mathcal{A}} \psi_k \le -(2\rho+\delta+\tau_\omega(s))]$ of Proposition 1, computed state-by-state from the measured budget. Only removing the exact monitor costs soundness, and it is the only variant whose CAR exceeds the sound ceiling.

Variant	CAR	$\mathrm{CAR}_{\mathrm{soft}}$	$\mathrm{CAR}_{\mathrm{th}}$	CBU	$\mathrm{CBU}_{\mathrm{all}}$	SR	Cost
CASTLE	0.381	0.950	0.353	0	0	0.89±0.11	0.12±0.03
w/o causal gate	0.284	0.284	0.131	0	0	0.90±0.05	0.13±0.05
legacy threshold	0.243	0.702	0.136	0	0	0.98±0.02	0.12±0.02
w/o exact monitor	0.950	0.950	0.761	3253	3253	0.89±0.11	0.12±0.03
w/o modulation	0.349	0.919	0.338	0	0	0.94±0.01	0.14±0.01
authority only	0.382	0.945	0.346	0	0	0.84±0.09	0.14±0.03
uniform LSE	0.215	0.215	0.003	0	0	0.97±0.02	0.11±0.01
PPO-Lag	0.271	0.271	0.011	0	0	0.97±0.02	0.09±0.04
product t-norm	0.658	0.658	0.009	975	923	0.99±0.01	0.10±0.03
Łukasiewicz	0.000	0.000	0.013	0	0	0.89±0.08	0.15±0.06

events $\{C_\omega \ge \rho+\delta\}$ and $\{\min_{k\in\mathcal{A}} m_k^\theta \ge \rho+\delta\}$ differ on a set of measure tending to zero and each bound converges to the corresponding exact-conjunction rate.

Lemma 2. Immediate from Definition 1; for (ii), $\psi_k = \psi_j - c$ implies $J_k = J_j$ exactly, so $\cos = 1$, and $m_k = m_j + c > m_j$ gives $j \prec k$ at every state.

C Exactness of the action Jacobians

The derivative $\partial\psi_k/\partial a$ is evaluated at the current action and the current state. For all three environments the one-step residual map is analytic and available in closed form, so J_k is exact up to floating point. When residuals are produced by a learned monitor, the gate inherits the monitor's Jacobian error; since ctrl is a ratio of norms, a multiplicative error on a single predicate does not bias the gate, while a systematic error on an agent-independent channel would. This is why we recommend auditing ctrl over a rollout window before removing a predicate from the attributable set.

Table 3: Per-predicate attribution on *SafeConf* (*CASTLE*, seed 0, averaged over all evaluation states). ctrl is authority, indep non-redundancy, cov the IARS gate; w^{assoc} is the conorm's own softmax weight and s^{causal} the same weight re-weighted by the gate and renormalised (Definition 1). The associational weight ranks the agent-independent monitor *first*; the gate ranks it last. AUROC against ground truth: 1.00 (gate), 0.67 (causal score), 0.53 (associational) and 1.00 (authority only).

#	predicate	role	gt	ctrl	indep	cov	w^{assoc}	s^{causal}
1	`speed`	attribute	1	1.00	0.80	0.80	0.047	0.219
2	`obst1`	attribute	1	1.00	0.74	0.74	0.080	0.288
3	`obst2(dup)`	redundant	0	1.00	0.00	0.00	0.015	0.000
4	`obst3(dup)`	redundant	0	1.00	0.00	0.00	0.009	0.000
5	`wall+x`	attribute	1	1.00	0.86	0.86	0.014	0.070
6	`wall-x`	attribute	1	1.00	0.89	0.89	0.017	0.078
7	`wall+y`	attribute	1	1.00	0.92	0.92	0.009	0.051
8	`exo-hazard`	exogenous	0	0.02	1.00	0.02	0.811	0.295

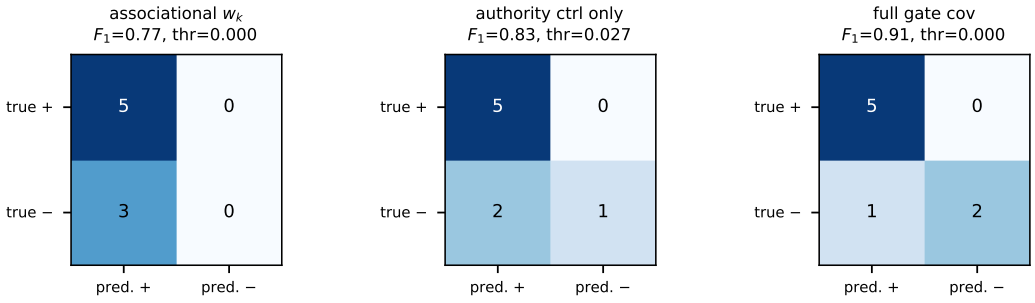


Figure 6: Attribution confusion matrices at the best-F_1 threshold, predicate level (5 positives, 3 negatives). Left: associational weights $w_k = \text{softmax}_k(-\beta m)$. Middle: authority only (ctrl). Right: full IARS gate. Only the full gate separates restatements from predicates with genuine authority.

D On the order-free variant

Replacing indep_k by the norm of J_k residualised against the *full* span $\text{span}\{J_j\}_{j\neq k}$ gives an order-free but rank-limited quantity: in a d-dimensional action space at most d predicates retain a non-zero residual, so with $K = 8$, $d = 2$ at most two predicates can be flagged as independent regardless of the specification. We measured the variant on *SafeConf* under the same trained policy (256 states per attribution step): the minimum-norm residual of J_k against the span of the other seven Jacobians is above 5% of $\|J_k\|$ for 0.01 of the eight predicates on average, and the resulting non-redundancy score is identically 1 for every predicate — the ranking carries no information and the gate degenerates to $P \equiv P_A$, which is the monolithic certificate we are trying to avoid. Two geometric predicates that are genuinely independent and lie in the shared two-dimensional image of the action map are equally indistinguishable. Conditioning on the causal order, as in Definition 1, removes the rank limitation at the cost of making the gate order-dependent; we make the order canonical via the margins, and the offsets in *SafeConf* make it state-independent. Under the same measurement the causal-order variant recovers $\text{indep} = 0.00$ for both restatements, 1.00 for the exogenous monitor and 0.69–0.87 for the five attributable predicates.

E What the attribution gate changes about the penalty

Because the budget is supported on $\mathcal{A}$, a predicate with $\text{cov}_k = 0$ contributes nothing to C_ω, hence nothing to the residual (2), hence nothing to the gradient $\partial c/\partial\theta = \sum_{k\in\mathcal{A}} w_k \partial m_k^\theta/\partial\theta$. In *SafeConf* the agent-independent monitor is violated on a large fraction of states, so a monolithic penalty spends a fixed fraction of its gradient budget on a term whose policy derivative is identically zero; empirically this inflates λ and over-constrains the controllable part. Table 1 shows the effect as

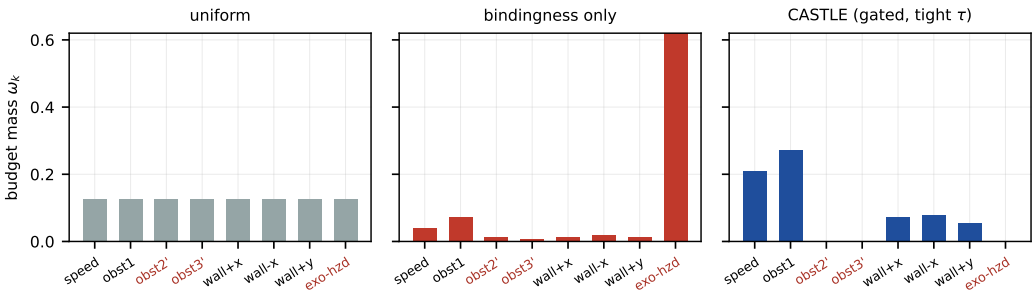


Figure 7: Budget mass per predicate. Uniform (left), bindingness-only (middle) and CASTLE (right). The gate moves mass off the agent-independent monitor and off the two restatements and onto predicates with independent authority, which is precisely what lowers τ_ω and raises the activation rate.

Table 4: Environment suite. All three are pure-PyTorch, vectorised, and expose analytic one-step residual maps, so action Jacobians are exact. Ground-truth control authority exists only in *SafeConf*, whose specification deliberately contains the two failure modes of associational attribution.

Env	K	$\lvert\mathcal{A}\rvert$	$\lvert S\rvert$	dynamics	specification note
SafeNavC	8	2	12	double integrator, 3 bouncing obstacles, 4 walls	clearance 0.30–0.50 to the goal
SafeReachC	6	2	8	2-link rigid-body arm, 4 substeps/step	joint/rate limits, obstacle, EE speed
SafeConf	8	2	12	SafeNavC physics, degraded specification	2 restatements + 1 exogenous monitor

a lower success rate at higher controllable cost for the uniform conorm relative to CASTLE on *SafeConf*, and no such gap on the two environments without a non-attributable predicate.

F Environments, hyper-parameters and cost

Shared learner. 16 parallel environments, rollout 128, 5 PPO epochs, minibatch 512, learning rate 5×10^{-4}, $\gamma = 0.99$, $\lambda_{\mathrm{GAE}} = 0.95$, clip 0.2, entropy bonus 0.01, value coefficient 1.0, gradient-norm clip 1.0, latent dimension 64, hidden width 128, dual learning rate 0.01, $\lambda_{\max} = 20$. Rewards and costs are standardised by a running estimate of their standard deviation so that the two critics remain fittable, a detail that turned out to be necessary for the LSE family and that we therefore apply to every method.[1]

Certificate parameters. $\rho = 0.05$, $\delta = 0.10$, $\delta_m = 0.02$, temperature initialised at $\beta_0 = 4$ and adapted by entropy feedback towards $H^\star = \log K/4$ clipped to $[0.5, 32]$, budget mix $\epsilon = 0.05$, gate threshold $\zeta = 0.5$, gate ramp $\mu = 0.20$, $\varsigma = 0.05$, redundancy band $(c_0, c_1) = (0.90, 0.995)$, attribution period $F = 4$ iterations over a batch of 512 states, gate EMA 0.8.

Attribution cost. Computing J_k for a batch requires one vectorised backward pass over the one-step residual map (K rows per state). On the CPU used for these experiments the extra time per attribution step is 3.2 ms for $K = 8$ at batch 512 (median of 30), against 0.47 s for a PPO iteration of 16×128 samples with 5 epochs; the amortised overhead at $F = 4$ is 0.2%. The causal-order

[1] One baseline detail is worth recording because it is easy to get wrong and it changes the comparison by a factor of five. A minimal-norm projection onto $\{\psi \le 0\}$ can leave the environment's action box; if the projected action is then clipped to the *environment* limits rather than to the policy's $\tanh$ support, the $\tanh^{-1}$ term in the log-probability saturates, the PPO ratio degenerates, and the baseline silently stops learning. Clamping every sweep to the policy's support and re-linearising at the projected point takes the safety layer from 0.09 to 0.53 success on *SafeNavC*. We report the corrected baseline everywhere.

variant, which sorts the predicates and residualises in sequence, costs $5.2\,\mathrm{ms}$ per batch, still below 0.3% amortised.