Universal Catastrophic Safety Undecidability and Capability–Risk Upper Bound Frontier: Unified Theorems, Complexity Positioning, and Engineering Pathways

Haobo Ma[1] Wenlin Zhang[2]
[1]Independent Researcher
[2]National University of Singapore

Abstract

Establish two foundational boundaries for general learning and decision systems. First, provide catastrophic safety determination undecidability for interactive agent–environment systems: under extremely weak modeling assumptions, for any extension-closed regular bad-prefix specification, whether threshold safety satisfied admits no global algorithm; under restricted subclass of deterministic environments and computable strategies, further position as Σ_1^0**-complete**/Π_1^0**-complete**. Second, provide **capability–worst-risk upper bound frontier** induced by joint PAC-Bayes high-probability bound, mutual information expected bound, and Wasserstein-1 distributionally robust optimization (Kantorovich–Rubinstein duality); via point perturbation adversary establish universal **geometric lower bound**, together with robust–accuracy impossibility and robust generalization sample complexity lower bound, forming "upper bound–lower bound" dual support. Thereby propose "scope restriction–runtime shielding–risk budget–structural prior" governance blueprint, providing minimal reproducible ImageNet-C + Shield experimental skeleton and metric configuration.

Keywords: Halting; Rice theorem; Σ_1^0-complete; POMDP; PAC-Bayes; conditional mutual information; Wasserstein-DRO; Kantorovich–Rubinstein duality; adversarial robustness; runtime shield; interruptibility

1 Introduction & Historical Context

Algorithmic decidability and program semantics reveal fundamental limits of universal verification: halting problem undecidable, Rice theorem states any non-trivial semantic property undecidable. Transplanting this idea to interactive agent–environment setting, obtain general determination unavailability for "whether triggering catastrophic specification". This direction resonates with undecidability results for infinite-horizon probabilistic planning/partially observable decision at threshold and plan existence. On the other hand, modern learning theory reveals capability and robustness cannot advance without cost: PAC-Bayes and mutual information paradigm characterize complexity/information amount influence on generalization, Wasserstein-DRO characterizes worst-risk under distribution shift; simultaneously, robust–accuracy impossibility and robust generalization sample complexity lower bound rigorously proven in natural model families. Two boundaries jointly point toward governance principles: acknowledging general static certification impossibility and capability–risk hard trade-off, adopt layered scheme of scope restriction, runtime shielding, and risk budget.

2 Model & Assumptions

2.1 Interaction Semantics and Temporal Assumptions

- Action and observation alphabets $\mathcal{A}, \mathcal{O}$ finite; history $h_{1:t} \in (\mathcal{A} \times \mathcal{O})^t$.
- **Computable policy**: Agent A is function $A : (\mathcal{A} \times \mathcal{O})^\star \to \mathcal{A}$, for any $h_{<t}$ exists finite time producing $a_t = A(h_{<t})$ at step t (allowing internal randomization via sampling program implementation). This assumption satisfied throughout undecidability construction and completeness positioning.
- Environment E specified by history conditional probability $\mu(o_t \mid h_{<t}, a_t)$. Main results use **deterministic trivial environment** E_0: always returns fixed observation $o_\perp$.

2.2 Safety Property and Regular Bad Prefix

- Let $\Sigma = (\mathcal{A} \times \mathcal{O})^\star$. Call $B \subseteq \Sigma$ **bad prefix language** if for any $u \in B$ and any extension $v \in \Sigma$, have $uv \in B$ (**extension-closed**). Corresponding **safe prefix set** $S = \Sigma \setminus B$ then **prefix-closed**.
- Specification adopts **regular bad prefix language** B (equivalent to safe prefix recognized by DFA/safety automaton). Violation event

$$\mathsf{Bad} = \{\exists t : \ h_{1:t} \in B\}, \qquad \tau(h) = \inf\{t : \ h_{1:t} \in B\}.$$

- **Threshold safety predicate**: Given $\varepsilon \in [0, 1)$,

$$\mathrm{Safe}_\varepsilon(A, E, B) := \Pr_\mu(\mathsf{Bad}) \ \le \ \varepsilon.$$

2.3 Learning–Evaluation and Distributional Robustness

- Data domain $\mathcal{Z}$ with metric d; sample $S = (Z_i)_{i=1}^n \sim D^n$.
- Learning algorithm outputs posterior $Q_S \in \mathcal{P}(\mathcal{H})$. Loss $\ell : \mathcal{H} \times \mathcal{Z} \to [0, 1]$.
- **Lipschitz assumption**: Exists uniform constant $L > 0$ such that for any $h \in \mathcal{H}$, mapping $z \mapsto \ell(h, z)$ is L-Lipschitz with respect to d (0-1 loss not applicable, adopt smooth surrogates like cross-entropy/hinge; controllable via spectral norm constraint and gradient clipping).
- Wasserstein-1 ball $\mathbb{B}_\rho(D) = \{D' : W_1(D', D) \le \rho\}$; robust risk

$$R_\rho^{\mathrm{rob}}(Q) := \sup_{D' \in \mathbb{B}_\rho(D)} \ \mathbb{E}_{h \sim Q, z \sim D'}[\ell(h, z)].$$

3 Main Results (Theorems and Alignments)

Theorem 1 (1: Universal Catastrophic Safety Determination Undecidable)**.** *Exists regular bad prefix family $\mathfrak{B}$ such that no algorithm can determine for all computable policies A, computable environments E, $B \in \mathfrak{B}$, and any rational $\varepsilon \in [0, 1)$ the truth value of* $\mathrm{Safe}_\varepsilon(A, E, B)$.

Theorem 2 (1': Complexity Positioning of Restricted Subclass)**.** *Under deterministic environment E_0 and computable policy class, let*

$$\mathit{UNSAFE} = \{(A, E_0, B, \varepsilon) : \Pr(\mathsf{Bad}) > \varepsilon\}, \quad \varepsilon < 1.$$

*Then UNSAFE is Σ_1^0-**complete**, its complement SAFE is Π_1^0-**complete**. In this subclass* $\Pr(\mathsf{Bad}) \in \{0, 1\}$*, thus "*$\Pr(\mathsf{Bad}) > \varepsilon$*" equivalent to "occurrence" for any $\varepsilon < 1$.*

Theorem 3 (2: Capability–Worst Risk Upper Bound Frontier: PAC-Bayes + KR)**.** *For any prior P and $\delta \in (0, 1)$, with probability at least $1 - \delta$ (over $S \sim D^n$) have*

$$\boxed{R_\rho^{\mathrm{rob}}(Q) \;\le\; \widehat{R}_S(Q) \;+\; \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln(1/\delta)}{2n}} \;+\; L\rho.}$$

*Right-hand three terms respectively empirical error, complexity/confidence term, and distribution shift linear penalty, constituting **upper bound induced capability–risk frontier**.*

Theorem 4 (2': High-Probability Mutual Information Bound: Paradigmatic Statement)**.** *Let loss $\ell \in [0, 1]$ with sub-Gaussian constant σ for each sample point. If learning algorithm satisfies **conditional mutual information** upper bound* $\mathrm{CMI}(S; Q_S) \le \Gamma$ *or equivalent strength uniform stability, then exists constant $c > 0$ such that for any $\delta \in (0, 1)$,*

$$\boxed{\Pr\left(R_D(Q_S) \;\le\; \widehat{R}_S(Q_S) \;+\; \sqrt{\frac{2\sigma^2(\Gamma + c\ln(1/\delta))}{n}} \right) \;\ge\; 1 - \delta.}$$

*Juxtaposing (2) with (1), obtain high-probability **frontier** expression for data-dependent posterior: take smaller of two right-hand sides as operational upper bound for capability–risk curve.*

Theorem 5 (3: Point Perturbation Geometric Lower Bound and Distribution Ball Inclusion)**.** *For any classifier f and $\rho > 0$, define*

$$\mathcal{R}_\rho^{\mathrm{adv}}(f) = \Pr_{(z,y)\sim D}\left[\exists z' \in B_\rho(z) :\; f(z') \ne y\right],$$

where $B_\rho(z) = \{z' : d(z, z') \le \rho\}$ with label preserving. Then

$$\boxed{\sup_{D' \in \mathbb{B}_\rho(D)} \quad R_{D'}(f) \;\ge\; \mathcal{R}_\rho^{\mathrm{adv}}(f).}$$

(3) holds on any metric and task, providing universal lower bound "foundation" matching (1). Under Gaussian mixtures and ℓ_p perturbations, exist constructive lower bounds for robust–accuracy impossibility and robust generalization sample complexity lower bound.

Proposition 6 (1: Tightness of KR Linear Term)**.** *For any $L, \rho > 0$ and metric space, exists L-Lipschitz function f and distribution pair (D, D') such that $W_1(D', D) = \rho$ and*

$$\sup_{W_1(D',D) \le \rho} \quad \mathbb{E}_{D'}[f] - \mathbb{E}_D[f] \;=\; L\rho.$$

Indicates first-order form $L\rho$ cannot be improved under uniform Lipschitz constant condition.

4 Proofs

4.1 Theorem 1 (Undecidability)

Take trivial environment E_0. Given Turing machine–input pair $\langle M, x\rangle$, construct computable policy

$$A_{M,x}(h_{<t}) = \begin{cases} a_\star, & \text{if } M(x) \text{ halts within } t \text{ steps,} \\ a_0, & \text{otherwise.} \end{cases}$$

Let bad prefix language $B = \{h : \text{ some step action is } a_\star\}$, regular and extension-closed. Then

$$\Pr(\mathsf{Bad}) = \mathbf{1}\{M(x) \text{ halts}\}.$$

If universal decider exists determining $\mathrm{Safe}_\varepsilon(A, E, B)$ truth/falsity for any input ($\varepsilon < 1$), obtain halting determination, contradiction. Proved.

4.2 Theorem 1' (Σ_1^0/Π_1^0 Complete)

Many-one reduction: Mapping $R : \langle M, x\rangle \mapsto (A_{M,x}, E_0, B, \varepsilon)$ polynomial-time computable, and $\langle M, x\rangle \in \text{HALT} \iff (A_{M,x}, E_0, B, \varepsilon) \in \text{UNSAFE}$ ($\varepsilon < 1$). **Membership**: Under E_0 and deterministic A, $\Pr(\mathsf{Bad}) \in \{0, 1\}$. If unsafe, exists minimal τ making $h_{1:\tau} \in B$, enumerate to this prefix accepts, thus UNSAFE $\in \Sigma_1^0$, complement in Π_1^0. Combining with reduction obtains completeness. Proved.

4.3 Theorem 2 (Upper Bound Frontier)

PAC-Bayes (McAllester/Catoni variant) provides

$$R_D(Q) \;\le\; \widehat{R}_S(Q) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln(1/\delta)}{2n}} \quad \text{(with probability} \ge 1 - \delta).$$

KR duality indicates for any L-Lipschitz function g,

$$\sup_{W_1(D', D) \le \rho} \mathbb{E}_{D'}[g] \;\le\; \mathbb{E}_D[g] + L\rho.$$

Applying to $g(z) = \mathbb{E}_{h\sim Q}\ell(h, z)$ yields (1). Proved.

4.4 Theorem 2' (High-Probability Mutual Information)

Let ℓ bounded with each point σ-sub-Gaussian. If algorithm satisfies $\mathrm{CMI}(S; Q_S) \le \Gamma$, then via information compression and variational inequality obtain

$$\Pr\!\left(R_D(Q_S) - \widehat{R}_S(Q_S) \;\le\; \sqrt{\tfrac{2\sigma^2(\Gamma + c\ln(1/\delta))}{n}} \right) \;\ge\; 1 - \delta,$$

where constant c given by tail control. Juxtaposing with (1) obtains frontier high-probability form. Proved.

4.5 Theorem 3 (Point Perturbation Lower Bound)

For any measurable selection operator $T : \mathcal{Z} \to \mathcal{Z}$ with $d(z, T(z)) \leq \rho$ almost surely, let $D' = (T, y)_{\#} D$. Taking coupling $\pi(dz, dz') = D(dz)\delta_{T(z)}(dz')$, then $\mathbb{E}_\pi d(Z, Z') \leq \rho$; thus $W_1(D', D) \leq \rho$. If $f(T(z)) \neq y$ then errs under D', further

$$\sup_{D' \in \mathbb{B}_\rho(D)} R_{D'}(f) \;\geq\; \mathbb{E}_D\big[\mathbf{1}\{\exists z' \in B_\rho(z) : f(z') \neq y\}\big] = \mathcal{R}_\rho^{\mathrm{adv}}(f).$$

Proved.

4.6 Proposition 1 (Tightness)

Take $D = \delta_0, D' = \delta_{\rho u}$ and $f(z) = L|z|_2$ yields result. Proved.

5 Model Apply

- **Autonomous control and tool-using agents**: Theorem 1 rules out general static certification, recommend restricting policy space and interfaces to verifiable sublanguages; during deployment suppress transgression via shields and interruptible protocols.
- **Perception–decision systems**: According to (1)(2) establish **risk budget**: under given (n, ρ) enhancing capability (larger model/weaker prior) requires correspondingly increased sample size, enhanced structural prior, or compressed L.
- **Evaluation and calibration**: Adopt corruption and perturbation benchmarks (e.g., ImageNet-C) and uncertainty measures (NLL/ECE), jointly "violation rate–task accuracy" dual-axis curves exhibiting "capability–risk frontier" and shield interception effectiveness.

6 Engineering Proposals

1. **Scope restriction**: Design policy and tool invocation via verifiable subsets (restricted DSL/interfaces), ensuring safety specifications implemented by online discrimination via DFA/LTL synthesis safe prefix recognizers.
2. **Runtime shield**: Synthesize pre-/post-shields via LTL→DFA→safety automaton generator; pre-shield filters unsafe action set, post-shield replaces with nearby safe action via minimal correction principle; probabilistic shield controls false rejection/false negative via confidence threshold.
3. **Risk budget**: Treat $(\widehat{R}_S, \mathrm{KL}, I, \rho, L)$ as budget quintuple; configure "data–prior–shift–Lipschitz" balancing strategy respectively during development and deployment phases.
4. **Structural prior and impact regularization**: Adopt equivariant structures, spectral norm constraints, and reversibility penalties (AUP) reducing complexity and side-effect propensity.
5. **Distributionally robust training and uncertainty governance**: Combine Wasserstein-DRO/adversarial training with deep ensembles, temperature calibration; handle high uncertainty via rejection–degradation–handoff open-loop strategy.
6. **Interruptibility**: Embed unbiased interruptible protocols in updating and exploration, preventing policy learning incentives to circumvent intervention.

7 Discussion (Risks, Boundaries, Past Work)

- **Boundary meaning**: Undecidability negates "general, global, one-time" static proof; under restricted model families (finite horizon, fully observable, discounted MDP, etc.) strong guarantees still obtainable.
- **Upper bound–lower bound enclosure**: KR linear term with PAC-Bayes/mutual information provide operational upper bounds; point perturbation lower bound with robust–accuracy impossibility, robust generalization sample complexity lower bound indicate "zero-cost both" unattainable not artifact of loose analysis.
- **Relationship with existing work**: Theorem 1 equivalent to Rice/halting, complements infinite-horizon probabilistic planning undecidability; Theorem 2 consistent with distributionally robust optimization, PAC-Bayes, mutual information paradigm; Theorem 3 matches constructive lower bounds and sample complexity lower bounds in adversarial robustness literature; shields and interruptibility correspond to runtime enforcement systems in safe reinforcement learning and formal methods.

8 Conclusion

General catastrophic safety determination unattainable in principle, capability enhancement and robustness admit hard trade-off jointly driven by complexity/information amount and distribution shift. Based on this, governance schemes should center on scope restriction, runtime shielding, and risk budget, proving within verifiable subdomain, backstopping via shields and interruptibility during deployment, suppressing shift risk via structural prior and distributionally robust techniques during training.

Acknowledgements, Code Availability

Thank related research in decidability, distributionally robust optimization, information-theoretic generalization, and safe reinforcement learning fields. Code and data not accompanying paper; minimal reproducible experiment suggestion: based on public corruption benchmarks, LTL→DFA tools, and adversarial/robust training libraries reproduce "risk budget curve–violation rate" dual-axis diagram; repository should include data scripts, specification examples, training/inference and shield modules, hyperparameter tables, and one-click scripts.

References

Turing, A. M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem.

Rice, H. G. (1953). Classes of Recursively Enumerable Sets and Their Decision Problems.

Madani, O., Hanks, S., Condon, A. (1999). On the Undecidability of Probabilistic Planning and Infinite-Horizon POMDPs.

McAllester, D. (1999). Some PAC-Bayesian Theorems.

Catoni, O. (2007). PAC-Bayesian Supervised Classification.

Alquier, P. (2021). User-friendly Introduction to PAC-Bayes Bounds.

Xu, A., Raginsky, M. (2017). Information-Theoretic Analysis of Generalization Capability of Learning Algorithms.

Steinke, T., Zakynthinou, L. (2020). Reasoning about Generalization via Conditional Mutual Information.

Bu, Y., Zou, S., Veeravalli, V. V. (2020). Tightening Mutual Information-Based Bounds on Generalization Error.

Villani, C. (2009). Optimal Transport: Old and New.

Esfahani, P. M., Kuhn, D. (2018). Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric.

Sinha, A., Namkoong, H., Duchi, J. (2018). Certifying Some Distributional Robustness with Principled Adversarial Training.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A. (2019). Robustness May Be at Odds with Accuracy.

Schmidt, L., et al. (2018). Adversarially Robust Generalization Requires More Data.

Alshiekh, M., et al. (2018). Safe Reinforcement Learning via Shielding.

Koenighofer, B., et al. (2024). Shields for Safe Reinforcement Learning.

Orseau, L., Armstrong, S. (2016). Safely Interruptible Agents.

Turner, A. M., Hadfield-Menell, D., Tadepalli, P. (2020). Conservative Agency via Attainable Utility Preservation.

Hendrycks, D., Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations.

A Semantics and Measurability

Let cylinder σ-algebra generated on Σ. Computable policy and environment jointly induce history distribution

$$\mathbb{P}(h_{1:t}) = \prod_{s=1}^{t} \big[A(a_s \mid h_{<s}) \cdot \mu(o_s \mid h_{<s}, a_s)\big].$$

Bad prefix language B extension-closed and regular, event $\mathsf{Bad} = \{\exists t: \ h_{1:t} \in B\}$ measurable; first violation time $\tau(h)$ is stopping time.

B "First Appearance $a_\star$" and Regular Bad Prefix

Define

$$B_{\text{hit}} = \{h: \ \exists i \leq |h|, \ a_i = a_\star\}.$$

If $u \in B_{\text{hit}}$ and v is any extension, then $uv \in B_{\text{hit}}$, thus extension-closed. Corresponding safe prefix set $S = \Sigma \setminus B_{\text{hit}}$ is prefix-closed.

C Binary Probability and Threshold Lemma

Under E_0 and deterministic A, Bad is event "whether appears $a_\star$", taking only values 0 or 1. For any rational $\varepsilon < 1$, have

$$\Pr(\mathsf{Bad}) > \varepsilon \iff \Pr(\mathsf{Bad}) = 1 \iff \mathsf{Bad} \text{ occurs.}$$

D Many-One Reduction Details

Mapping R sends $\langle M, x\rangle$ to $(A_{M,x}, E_0, B_{\text{hit}}, \varepsilon)$.

- **Correctness**: If $M(x)$ halts, exists t_0 making $A_{M,x}$ output $a_\star$ at t_0, thus Bad occurs; otherwise not.
- **Computability**: Constructing $A_{M,x}$ and DFA recognition for B_{hit} both completed in polynomial time.
- **Completeness**: By HALT $\leq_m$ UNSAFE and membership obtain Σ_1^0-complete; complement problem obtains Π_1^0-complete.

E PAC-Bayes and KR Duality Composition

Let $g_Q(z) = \mathbb{E}_{h\sim Q}\ell(h,z)$. If $\ell \in [0,1]$ and $z \mapsto \ell(h,z)$ uniformly L-Lipschitz, then g_Q also L-Lipschitz. KR duality provides

$$\sup_{W_1(D',D)\leq\rho} \mathbb{E}_{D'} g_Q \leq \mathbb{E}_D g_Q + L\rho.$$

PAC-Bayes basic formula bounds $\mathbb{E}_D g_Q$ and $\widehat{R}_S(Q)$ difference with probability $1-\delta$, composition yields (1).

F Mutual Information High-Probability Bound (CMI Paradigm)

Under $\ell \in [0,1]$ and pointwise σ-sub-Gaussian, conditional mutual information $\mathrm{CMI}(S; Q_S) \leq \Gamma$ induces

$$\Pr\left(R_D(Q_S) - \widehat{R}_S(Q_S) \leq \sqrt{\tfrac{2\sigma^2(\Gamma + c\ln(1/\delta))}{n}} \right) \geq 1-\delta.$$

Proof based on information compression inequality and PAC-Bayesian-style variational techniques; when replacing CMI with uniform stability, same-order tail bound obtainable.

G Point Perturbation Lower Bound Measurable Selection and Label Preserving

Let metric space separable with complete Borel σ-algebra. When selecting for each (z,y) $z' \in B_\rho(z)$ making f err, adopt Borel measurable selection lemma defining operator $T(z)$; label preserving assumption ensures pushforward $D' = (T,y)_\# D$ consistent with task. If error-causing perturbation non-existent, set $T(z) = z$. This yields (3).

H Lipschitz Constant and Surrogate Loss

0-1 loss does not satisfy Lipschitz assumption; use surrogate losses like cross-entropy/hinge, control L via spectral norm constraints, gradient clipping, and Lipschitz network structures. This control enters (1) linear term, determining ρ-sensitivity.

I Engineering Metrics and Indicators

Risk budget curve: Horizontal axis is complexity/information term (model scale or prior strength, $I(S;Q_S)$ proxy), vertical axis is R_ρ^{rob} estimate or its upper bound; overlay "violation rate–task accuracy" dual-axis curves (with/without shield two curves), exhibiting runtime shielding violation suppression effect at similar accuracy.