Uncertainty-Aware Policy-Preserving Abstractions with Abstention for One-Shot Decisions

Sandy Tanwisuth*
Center for Human-Compatible AI (CHAI), UC Berkeley
ML Alignment & Theory Scholars (MATS) Program
kst@berkeley.edu

Daniel K. Leja
Independent
danielleja3@gmail.com

Abstract

State abstractions in reinforcement learning traditionally preserve optimal policies but discard information about decision confidence, preventing principled abstention when choices are uncertain. We propose that optimization—the dominant paradigm in RL—may itself be an inadequate lens for safety-critical deployments where the challenge is not achieving optimal performance, but recognizing boundaries where confident action risks unsafe amplification of uncertainty. We advocate for uncertainty-aware, policy-preserving abstractions: represent states not just by which actions are optimal, but by how strongly they dominate alternatives—their decision margins. When margins are small, systems should abstain rather than commit to weakly-supported choices. This integrates abstention directly into the state representation rather than adding confidence checks as an afterthought. This reframing reveals three fundamental challenges that resist traditional optimization: computational—when can we efficiently compute margins in large action spaces, and what structure makes this tractable? statistical—how do we learn which states require which margins without access to true rewards or deferral costs? epistemic—how do we set abstention thresholds when the costs of errors versus deferrals are fundamentally uncertain or contested? We argue these aren't merely technical gaps but potential boundaries where optimization itself becomes problematic. This position paper formalizes the margin-based framework and invites both theorists and experimentalists to investigate whether these challenges represent surmountable technical barriers or fundamental limits requiring new theoretical foundations beyond optimization.

1 Introduction

Autonomous systems increasingly face scenarios where recognizing when *not* to act is as crucial as selecting optimal actions. Medical diagnosis systems must identify cases requiring specialist review [Rajpurkar et al., 2022, Topol, 2019]; autonomous vehicles must recognize situations beyond their operational design domain [Koopman and Wagner, 2016, Shalev-Shwartz et al., 2016]; automated trading systems must detect anomalous market conditions [Kirilenko et al., 2017]. Despite this ubiquity, current abstraction theory—largely rooted in optimization paradigms—provides limited guidance on principled abstention mechanisms.

Classical state abstraction theory groups contexts that share optimal actions, enabling efficient decision-making on simplified representations [Li et al., 2006, Givan et al., 2003, Abel et al., 2018]. While these approaches elegantly compress state spaces when clear optima exist, they may become problematic near decision boundaries where small perturbations significantly change outcomes, potentially amplifying rather than mitigating errors.

*Corresponding author: kst@berkeley.edu

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

Traditional approaches to uncertainty in RL—including safe reinforcement learning [García and Fernández, 2015, Amodei et al., 2016], conformal prediction [Angelopoulos and Bates, 2021], and learning to defer [Madras et al., 2018, Mozannar and Sontag, 2020]—address uncertainty within optimization frameworks: minimizing regret, maximizing reward subject to constraints, or optimizing cost-sensitive objectives that trade errors against deferrals. While these methods add safety mechanisms to optimization, they don't question whether optimization itself might be inappropriate in regions of high uncertainty.

Consider a medical diagnosis system at the boundary between two treatments with nearly identical expected utilities. An optimization-driven approach would confidently choose the marginally better option. Yet this confidence might represent epistemic overreach—the system may be unable to distinguish between genuine optimality and noise in finite samples. Perhaps the issue is not finding the optimal action, but recognizing where the concept of optimality itself becomes unreliable.

This raises a foundational question: *Can we develop abstractions that simultaneously compress state spaces and preserve the information necessary for principled abstention?* We explore this through margin-based abstractions—augmenting traditional abstractions with decision confidence measures. This analysis reveals three interconnected open problems whose resolution may require moving beyond an optimization-centric framework.

Position. We contend that forcing decisions through optimization—always selecting the action with highest expected value—becomes problematic near decision boundaries where margins are small and uncertainty is high. Rather than treating abstention as an afterthought, we propose building it directly into state representations through *margin-based abstractions*: augment policy-preserving abstractions with explicit measures of how strongly optimal actions dominate alternatives. This enables principled act-or-defer decisions: commit only when margins exceed justified thresholds, whether derived from risk-cost trade-offs (act when estimated regret is below deferral cost) or calibrated coverage (act when prediction sets single out one action). While conceptually simple, this approach reveals fundamental challenges: computing margins may be intractable without exploitable structure, learning which margins matter requires error costs that may not exist in meaningful form, and setting thresholds challenges core assumptions when objectives are contested or ill-defined. We argue these challenges suggest optimization may be the wrong framework entirely in regions of high uncertainty—not due to algorithmic limitations, but because optimality itself becomes unreliable. This paper formalizes the margin-based framework and articulates why these boundaries may require theoretical foundations that acknowledge rather than optimize away epistemic uncertainty.

2 Margin-Based Abstractions

We formalize our framework in the contextual bandit setting as a foundation for broader application to RL. Consider a context space $\mathcal{X}$ with distribution $\mathcal{D}$ and finite action set $\mathcal{A} = \{a_1, \ldots, a_m\}$. Choosing action $a \in \mathcal{A}$ in context $x \in \mathcal{X}$ yields bounded reward $R \in [0, 1]$. Define the expected utility $U(x, a) := \mathbb{E}[R \mid X = x, A = a]$ and optimal action set $B(x) := \arg\max_{a \in \mathcal{A}} U(x, a)$.

2.1 Augmenting Abstractions with Decision Margins

An *abstraction* is a mapping $\phi : \mathcal{X} \to \bar{\mathcal{X}}$ that partitions contexts into equivalence classes. Standard approaches preserve only optimal actions:

Definition 1 (Policy-Preserving Abstraction)**.** *An abstraction $\phi : \mathcal{X} \to \bar{\mathcal{X}}$ is policy-preserving if $x \sim_\phi x' \implies B(x) = B(x')$, where $x \sim_\phi x'$ denotes $\phi(x) = \phi(x')$.*

This ensures contexts grouped together require the same optimal actions. For each abstract state $\bar{s} \in \bar{\mathcal{X}}$, define:

$$U_\phi(\bar{s}, a) := \mathbb{E}[U(X, a) \mid \phi(X) = \bar{s}] \tag{1}$$

$$B_\phi(\bar{s}) := \arg\max_{a \in \mathcal{A}} U_\phi(\bar{s}, a) \tag{2}$$

Our key innovation is augmenting these abstractions with explicit measures of decision confidence:

Definition 2 (Decision Margin)**.** *For abstract state $\bar{s}$ with optimal utility $M_1(\bar{s}) := \max_a U_\phi(\bar{s}, a)$ and runner-up utility $M_2(\bar{s}) := \max_{a \notin B_\phi(\bar{s})} U_\phi(\bar{s}, a)$, the decision margin is:*

$$\gamma_\phi(\bar{s}) := M_1(\bar{s}) - M_2(\bar{s})$$

When all actions are equally optimal, we define $M_2(\bar{s}) = M_1(\bar{s})$ to ensure $\gamma_\phi(\bar{s}) = 0$.

The margin $\gamma_\phi(\bar{s})$ serves not as a measure of "how optimal" a decision is—which presupposes optimization as the goal—but as an indicator of *epistemic boundaries*: regions where the abstraction can confidently act (large γ) versus where it may struggle to distinguish signal from noise (small γ). This reframing shifts focus from maximizing expected utility to recognizing where the concept of optimality itself becomes unreliable.

2.2 Abstention as Epistemic Caution

Given an abstention threshold $\theta(\bar{s})$, the policy becomes:

$$\pi_\phi(\bar{s}) = \begin{cases} a^* \in B_\phi(\bar{s}) & \text{if } \gamma_\phi(\bar{s}) \geq \theta(\bar{s}) \\ \perp & \text{otherwise} \end{cases}$$

where $\perp$ represents deferral to external decision-making (e.g., human judgment).

Our framework differs from existing confidence-based approaches in a crucial way: the margin $\gamma_\phi(\bar{s})$ emerges directly from the abstraction's partition structure rather than requiring an auxiliary uncertainty estimator applied post-hoc. This integrates abstention into the representation itself, enabling simultaneous state space compression and uncertainty-aware decision-making. However, this integration creates new theoretical challenges: the abstraction must be learned without knowing which margins matter, margins must be estimated without ground truth, and thresholds must be set without clear objectives. These interdependencies, explored in the next section, suggest that margin-based abstractions may require fundamentally different theoretical foundations than traditional optimization frameworks provide.

3 Three Open Problems

We now present three interconnected problems that margin-based abstractions reveal. These are not merely technical challenges awaiting better algorithms, but potential boundaries where optimization itself becomes an inadequate framework for decision-making.

3.1 Problem 1: Can Structure Break the Curse of Actions?

Computing margins appears to require comparing all actions, potentially scaling linearly with $|\mathcal{A}|$. Yet empirical successes in contextual bandits that exploit action structure [Hanna et al., 2023, Zhu et al., 2022] and hierarchical decompositions in RL [Kulkarni et al., 2016, Nachum et al., 2018] suggest that real-world problems contain structure that could enable more efficient margin computation.

Open Problem 1 (The Structure-Complexity Gap)**.** *Characterize conditions under which margin detection complexity can be substantially reduced by exploiting problem structure.*

Consider candidate structures that might enable improved scaling:

- **Sparsity:** Only $k \ll |\mathcal{A}|$ actions are ever near-optimal
- **Metric structure:** Actions embed in low-dimensional spaces with smooth utilities
- **Hierarchical decomposition:** Actions organize into trees or taxonomies

Open questions: Can structural assumptions reduce margin detection from $O(|\mathcal{A}|)$ comparisons to polynomial dependence on structural parameters (e.g., $O(k \log |\mathcal{A}|)$ for sparsity)? How do we identify which structure applies without exhaustive exploration? Are there fundamental information-theoretic lower bounds on margin detection regardless of structure?

Why this matters: Medical diagnosis systems confront thousands of potential treatments; autonomous vehicles face exponentially many trajectory options. Without exploiting structure, margin

computation remains intractable. Work on contextual bandits with large action spaces [Hanna et al., 2023, Zhu et al., 2022] demonstrates that structure enables practical algorithms, but these don't directly address margin preservation—the gap between their success and our requirements needs bridging.

3.2 Problem 2: The Bootstrap Paradox

Learning margin-preserving abstractions requires knowing which contexts can be grouped together—but determining valid groupings requires knowing the margins. This circular dependency challenges the feasibility of the entire approach.

Open Problem 2 (Learning Without Ground Truth)**.** *Design algorithms that learn margin-preserving abstractions from finite samples without oracle access to the correct abstraction.*

The fundamental tensions creating this paradox:

- Estimating margins reliably requires sufficient samples per abstract state
- Getting sufficient samples requires coarse abstractions with few states
- Coarse abstractions may incorrectly merge contexts with different margins
- Detecting incorrect merging requires reliable margin estimates

Open questions: Can pre-trained representations break this circularity by providing initial structure? How do we adapt foundation models—trained for prediction, not decision-making—to preserve decision margins? What theoretical guarantees can we provide when bootstrapping from pre-trained features versus learning from scratch?

Why this matters: The success of foundation models suggests a path forward: leverage pre-trained representations as scaffolding for margin-preserving abstractions. This bridges theory and practice—theorists need to characterize when pre-training provides sufficient structure, while practitioners need methods to adapt existing models for abstention-aware decision-making. The gap between pre-training objectives (next-token prediction) and decision requirements (margin preservation) defines a critical research frontier.

3.3 Problem 3: Abstention Without Known Costs

Even with perfect abstractions and margins, setting abstention thresholds traditionally requires numerical costs for errors versus deferrals. But these costs often cannot be meaningfully quantified—not because they are unknown, but because they may not exist as coherent numerical values.

Open Problem 3 (The Unobservable Cost Problem)**.** *Develop principled methods for setting abstention thresholds without reducing decisions to numerical cost-benefit trade-offs.*

The fundamental challenge: traditional optimization assumes commensurable costs, but real abstention decisions involve:

- Binary feedback (human accepts/rejects) rather than numerical penalties
- Incommensurable values (preventing harm vs. providing assistance)
- Context-dependent boundaries that shift with social norms and stakeholder values
- Ethical considerations that resist quantification

Open questions: Can we reframe abstention as learning decision boundaries from human interventions rather than optimizing cost trade-offs? When humans consistently override system decisions in certain contexts, does this reveal learnable abstention boundaries independent of explicit costs? How can theory explain the empirical success of learning boundaries from feedback?

Why this matters: Modern RLHF systems already learn abstention boundaries without explicit costs—language models learn when to refuse harmful requests through human feedback, not cost functions. Content moderation systems develop abstention patterns from accepted/rejected examples. This empirical success suggests abstention boundaries can be treated as objects to be learned from revealed preferences rather than derived from optimization. Understanding when and why this works—and when it fails—is critical for safe deployment.

Acknowledgments and Disclosure of Funding

This work was co-conceptualized by Sandy Tanwisuth and Daniel K. Leja. Sandy led the research development, theoretical framing, and synthesis, including engagement with mentors and colleagues; Daniel contributed to the direction, articulation, refinement, and coordination of the project. Sandy thanks Niklas Lauffer for mentorship during the CHAI internship, and Richard Ngo for mentorship as part of the Coalitional Agency research agenda during MATS 8.0. Sandy also thanks Jeffrey Hennigan and the MATS Program for logistical support, and Michael Jemison and Aaron Kirtland for insightful discussions. This work was conducted by Sandy Tanwisuth during MATS 8.0, building on ideas developed during a CHAI internship, and was supported by external funding from the Cooperative AI Foundation Early Career Researcher.

References

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *ICML*, pages 10–19, 2018.

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In *NeurIPS*, pages 9558–9570, 2019.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv:1606.06565*, 2016.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *ICML*, pages 136–145, 2017.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*, 2021.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *FOCS*, pages 207–216, 2013.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv:1306.6709*, 2013.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021.

Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Information Theory*, 52(12):5406–5425, 2006.

Clément L. Canonne. A Survey on Distribution Testing: Your Data is Big. But is it Blue? *Theory of Computing Library Graduate Surveys*, 9:1–100, 2020.

Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. MICo: Improved representations via sampling-based state similarity for Markov decision processes. In *NeurIPS*, 2020.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *arXiv preprint arXiv:1706.03741*, 2023. URL: `https://arxiv.org/abs/1706.03741`.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

David L. Donoho. Compressed sensing. *IEEE Trans. Information Theory*, 52(4):1289–1306, 2006.

Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, pages 973–978, 2001.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017.

Dylan J. Foster, Akshay Krishnamurthy, and Haipeng Luo. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *NeurIPS*, 2021.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *JMLR*, 16(1):1437–1480, 2015.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *ALT*, pages 174–188, 2011.

Itzhak Gilboa. *Theory of Decision under Uncertainty*. Cambridge University Press, 2009.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.

Dennis Gross, Nils Jansen, and Guillermo Perez. Robustness verification for classifier ensembles. In *ATVA*, pages 271–287, 2020.

Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *NeurIPS*, pages 3909–3917, 2016.

Josiah Hanna, Shengjie Yang, and Christina Fragouli. Efficient batched algorithm for contextual linear bandits with large action space via soft elimination. In *NeurIPS*, 2023.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE TPAMI*, 44(9):5149–5169, 2021.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Information Theory*, 61(5):2835–2885, 2015.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv:2207.05221*, 2022.

Andrei A. Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *Journal of Finance*, 72(3):967–998, 2017.

Philip Koopman and Michael Wagner. Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety*, 4(1):15–24, 2016.

Tejas D. Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *NeurIPS*, pages 3675–3683, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *International Symposium on AI and Mathematics*, 2006.

Marlos C. Machado, Marc G. Bellemare, and Michael Bowling. A Laplacian framework for option discovery in reinforcement learning. In *ICML*, pages 2295–2304, 2017.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *NeurIPS*, pages 6147–6157, 2018.

Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *JMLR*, 8:2169–2231, 2007.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *ICML*, pages 7076–7087, 2020.

Rémi Munos. From bandits to Monte-Carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–129, 2014.

Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. In *NeurIPS*, pages 3303–3313, 2018.

Gergely Neu and Gábor Bartók. An efficient algorithm for learning with semi-bandit feedback. In *ALT*, pages 234–261, 2020.

Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *ICML*, pages 663–670, 2000.

Andi Peng, Aviv Netanyahu, Mark Ho, et al. Learning with language-guided state abstractions. In *ICLR*, 2024.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, et al. Deep reinforcement learning for sepsis treatment. *Nature Medicine*, 25(11):1795–1803, 2019.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv:1908.05659*, 2019.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. AI in health and medicine. *Nature Medicine*, 28(1):31–38, 2022.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *JAIR*, 48:67–113, 2013.

Leonard J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46(253):55–67, 1951.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv:1610.03295*, 2016.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, pages 1–5, 2015.

Eric J. Topol. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

Zizhao Wang, Xuesu Xiao, Zifan Xu, et al. CARL: A benchmark for causal abstraction in vision-based reinforcement learning. In *NeurIPS*, 2024.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. Technical Report HPL-2003-97R1, HP Labs, 2003.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *ICLR*, 2021.

Yinglun Zhu, Dylan J. Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *ICML*, 2022.

A Motivating Example

Consider a diagnostic system with context space $\mathcal{X}$ (patient features) and action space $\mathcal{A}$ (treatment recommendations). A standard abstraction might group patients with similar optimal treatments, but this discards critical information:

Two patient groups might both have "prescribe antibiotic A" as optimal:

- Group 1: $U(\text{antibiotic A}) = 0.9$, $U(\text{antibiotic B}) = 0.3$ yields $\gamma = 0.6$ (confident)
- Group 2: $U(\text{antibiotic A}) = 0.51$, $U(\text{antibiotic B}) = 0.49$ yields $\gamma = 0.02$ (uncertain)

Standard abstractions treat these identically, but Group 2 warrants specialist consultation. Our margin-based approach preserves this distinction, enabling principled deferral when $\gamma < \theta$. This illustrates why preserving margins is orthogonal to existing approaches: conformal prediction provides confidence intervals on individual predictions but doesn't address state space reduction, while safe RL focuses on avoiding catastrophic actions rather than identifying decision uncertainty.

A Technical Details and Challenges

This appendix provides extended technical discussion of the challenges identified in Section 3. These details illuminate why the three problems may require fundamentally new theoretical approaches rather than extensions of existing methods.

A.1 Problem 1: Computational Complexity Challenges

The margin detection problem faces several interrelated technical barriers that existing theory does not adequately address:

Instance-dependent bounds for contextual bandits [Foster et al., 2021, Lattimore and Szepesvári, 2020] provide improved sample complexity under structural assumptions, but require knowing the structure class a priori. In practice, we must simultaneously discover which structure applies and exploit it for efficient margin computation—a joint learning problem that lacks theoretical characterization.

More fundamentally, existing work optimizes for identifying single best actions, while margin preservation requires maintaining information about all near-optimal actions and their relative rankings. This shifts the problem from point estimation to preserving partial orders over action sets, potentially requiring sample complexity that scales with the number of near-optimal actions rather than just the action space size. The interaction between abstraction granularity (how coarsely we partition states), margin fidelity (how accurately we preserve decision boundaries), and sample efficiency remains an open theoretical question.

A.2 Problem 2: Representation Learning Challenges

The bootstrap paradox extends beyond a simple circular dependency to reveal fundamental gaps in our understanding of representation learning for decision-making:

Bisimulation metrics [Castro et al., 2020] preserve value equivalence but can incorrectly group states with vastly different margins—two states might have $Q(s_1, a) \approx Q(s_2, a)$ for all actions yet have margins $\gamma(s_1) \gg \gamma(s_2)$. This occurs because bisimulation focuses on preserving value functions rather than decision boundaries.

Recent approaches using language-guided [Peng et al., 2024] or causal abstractions [Wang et al., 2024] sidestep the bootstrap problem by importing external structure, but provide no theory for when such structure suffices for margin preservation. The gap between pre-training objectives (prediction, reconstruction) and decision requirements (margin preservation) remains unbridged. We need new theory connecting representation quality to decision-theoretic properties—specifically, what properties must representations satisfy to enable learning margin-preserving abstractions without ground truth?

A.3 Problem 3: Decision-Theoretic Challenges

The absence of meaningful cost functions challenges the foundations of decision theory:

Traditional frameworks—inverse RL [Ng and Russell, 2000], cost-sensitive learning [Elkan, 2001], preference-based methods [Christiano et al., 2023]—all assume that preferences can ultimately be reduced to numerical utilities, even if these are initially hidden or uncertain.

However, empirical success in RLHF and content moderation demonstrates that systems can learn appropriate abstention boundaries from binary feedback alone, without ever specifying or learning explicit costs. This suggests abstention boundaries might be primary objects that emerge from human-AI interaction patterns rather than derived consequences of optimization. Margin-based abstractions provide a framework for studying these boundaries directly, but we lack the mathematical tools to characterize when and why learning from revealed boundaries succeeds without underlying cost functions.

B Conclusion

Margin-based abstractions offer a framework for studying these questions by making abstention boundaries explicit objects of analysis rather than implicit consequences of optimization. Whether the challenges we identify represent surmountable technical barriers or fundamental limits of optimization-based decision-making remains an open question—one that will require both theoretical innovation and empirical investigation to resolve.

The three problems we identify—computational tractability of margin detection, learning abstractions without ground truth, and setting thresholds without costs—reveal potential boundaries where optimization itself becomes problematic, not merely difficult. These interdependencies suggest that the dominant paradigm in RL may be inadequate for safety-critical deployments where recognizing epistemic boundaries matters more than achieving optimal performance. The challenge is not optimizing better, but acknowledging regions where confident action risks unsafe amplification of uncertainty. This position paper invites both theorists and experimentalists to investigate whether these represent temporary technical gaps or fundamental limits requiring new theoretical foundations beyond optimization.