Infra-Bayesian Reinforcement Learning Agents Outperform Classical RL For Worst-Case Robustness

ARXIV CODE

Accepted to the Finding the Frame Workshop at RLC 2026

[*]Manish Aryal[1] · [*]Faiyaz Azam[2] · [*]Agnivo Banerjee[3] · [*]Syed Mahir Ahamed[4] · [*]Sai Sidhanth Manoharan Jayanthi[5] · [*]Allegra Laro[6] · [*]Clément Legentilhomme[7] · [*]Andrew Lin[8] · [*]Florian Lorkowski[9] · [*]Marina Pérez del Valle[10] · [*]Radman Rakhshandehroo[11] · [*]Patric Rommel[12] · [*]Emanuel Ruzak[13] · [*]Nathan Theng[14] · [†]Paul Yushin Rapoport[15]

[1]Purdue University · [2]Carnegie Mellon University · [3]WorldQuant University · [4]Union College · [5]UC Berkeley · [6]Independent · [7]Aix-Marseille University · [8]MIT · [9]University of Zurich · [10]University of Massachusetts Amherst · [11]University of British Columbia · [12]University of Stuttgart · [13]University of Buenos Aires · [14]California State University, Fresno · [15]University of Chicago

Co-first author, contributed equally, order alphabetical.

†Mentor and corresponding author. *Kairos and the SPAR program provided funding and logistical support.*

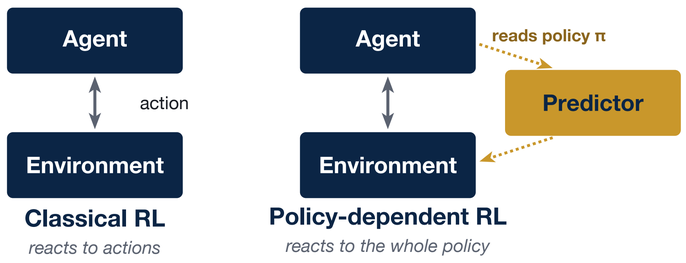
Infra-Bayesianism · Robust RL · Knightian Uncertainty · Newcomb's Problem

Abstract

Classical RL assumes a fixed environment that never reacts to the agent's own policy, an assumption that breaks down once predictors, markets, or institutions anticipate the agent's behavior. Under model misspecification, Bayesian RL agents can become **confidently wrong**. Infra-Bayesianism (IB) instead scores each policy by its **worst case** across a set of admissible hypotheses. We built the **first proof-of-concept IB reinforcement learning agent for finite-outcome, stateless decision problems** and tested it under Knightian uncertainty and in Newcomb's problem: it matches classical Bayesian RL with no ambiguity, cuts worst-case regret sharply when the environment is genuinely uncertain, and picks the **optimal strategy** in Newcomb's problem, where classical decision theories disagree with each other.

Why Classical RL Breaks

RL is analyzed as if the environment were fixed, but predictors, markets, and institutions model the agent and respond to its **whole policy**, not just its actions [1].



Bayesian RL also needs realizability, and no tractable hypothesis class contains a world that includes the agent itself [2].

Infra-Bayesianism replaces the single prior with a **set** of admissible hypotheses and scores each policy by its **worst case**, so Knightian uncertainty is never quietly averaged away.

How This Fits In

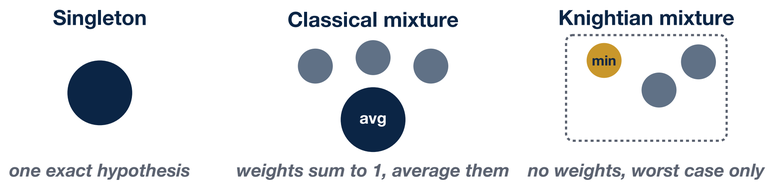
Three nearby literatures each solve part of the problem. Value-based RL agents can fail to converge under policy-dependence [1], and none of the three combines worst-case evaluation, set-based uncertainty, policy-dependence, and IB's offset term in one framework.

	Worst-case value	Set-based unc.	Policy-dep.	Offset term
Robust RL	✓	✓	×	×
Credal sets	✓	✓	×	×
Policy-dependent RL	×	×	✓	×
This work	✓	✓	✓	✓

Institutions

Method: Combining Hypotheses

A hypothesis is an **a-measure** $a = (\lambda\mu, b)$, a measure μ scaled by λ plus an offset b holding the value of branches already ruled out [3]. The agent's **hypothesis set** is a collection of these, scored by its worst member, $\inf_{a\in\Psi} a(f)$. Sets arise three ways:



Method: The Decision Loop

Every step, the agent scores each policy by its worst case and acts on the best guarantee (**maximin**), then updates by the IB conditioning rule.

Figure 1. The agent's decision loop: evaluate every policy by its worst case, act on the best guarantee, then fold the observation back into belief state.

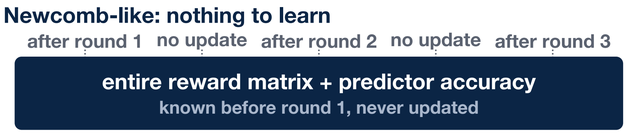
Method: Is There Anything to Learn?

Each environment gets its own **world model**: the rules for representing measures and deriving predictions. The two below differ on whether there is anything to learn.

Bernoulli bandit: state accumulates

learning = updating the (N, R) counts, which recreates Bayes' rule

Newcomb-like: nothing to learn



the structure is known from the start, the agent computes the answer once

Classical-recovery validation uses the first world model, Newcomb's Problem uses the second; the trap bandit experiment uses a separate three-outcome joint model, not pictured here.

Our Contributions

- **Finite-outcome IB architecture.** a-measures, infradistributions, classical/Knightian mixtures, IB updates.
- **Bayes as a special case.** Exact Bayesian behavior when all uncertainty is classical.
- **Lower worst-case regret under Knightian uncertainty.** 29.0 vs. 52.0 at t=100.
- **Optimal in Newcomb's problem.** Classical decision theories disagree on it.

Result: Newcomb's Problem

Payoff	Pred. one-box	Pred. two-box
One-box	**$10**	$0
Two-box	**$11**	$1

In Newcomb's problem [4], one-boxing pays **$10** against a good predictor, where two-boxing pays **$1**. The IB agent switches at exactly the right point (**α above 0.55**), while a causal-decision agent two-boxes regardless.

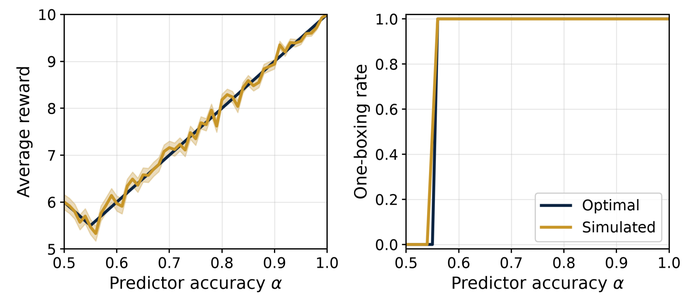


Figure 2. The IB agent consistently selects the optimal Newcomb's-problem policy across predictor accuracy α, over 1000 simulated episodes.

Result: Robust to Misspecification

With a severely misspecified prior ($\alpha_{\text{prior}} = 0.01$ against a true risky rate of 0.99), the Bayesian agent keeps pulling the trapped arm: **609.6** median regret, **65%** catastrophe rate. The IB agent treats safe-vs-risky as Knightian instead of committing to a point prior, holding regret to **9.6** and catastrophes to **4%**, matching what a correctly specified prior would have given. This has a cost: in a mostly-safe world (α=0.01), IB regret is **39.6** against **0.4** for a correctly specified Bayesian agent.

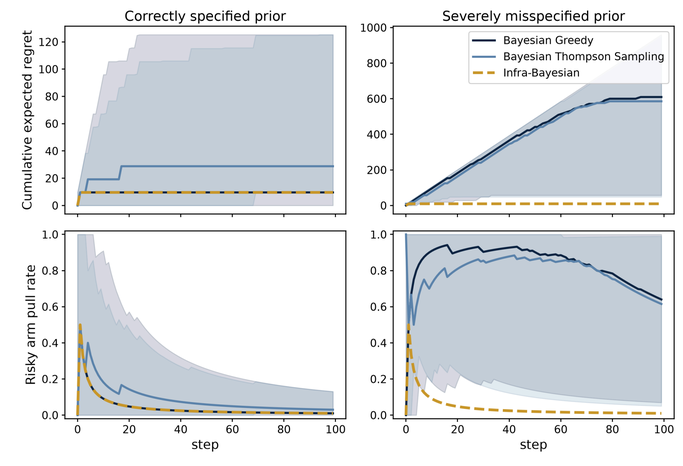


Figure 3. A misspecified Bayesian prior drives regret to 609.6 and a 65% catastrophe rate (right); the dashed gold Infra-Bayesian line holds regret to 9.6 and catastrophes to 4% either way.

Conclusion

Infra-Bayesian reasoning works as an **actual RL agent**, not just a theoretical framework: it recovers ordinary Bayesian behavior as a special case, and separating probabilistic from Knightian uncertainty produces measurably more robust behavior when that distinction matters. The current implementation is limited to finite outcomes and small hypothesis spaces; **continuous states and multi-step decisions** are ongoing work toward RL that is robust by design.

References

[1] J. Bell, L. Linsefors, C. Oesterheld, and J. Skalse, "Reinforcement Learning in Newcomblike Environments," in *Advances in Neural Information Processing Systems*, 2021, pp. 27284–27295.

[2] B. Gelb, "What is Inadequate about Bayesianism for AI Alignment: Motivating Infra-Bayesianism," *AI Alignment Forum*, May 2025, [Online]. Available: https://www.alignmentforum.org/posts/wzCtwYtojMabyEg2L/what-is-inadequate-about-bayesianism-for-ai-alignment

[3] A. Appel, "Basic Inframeasure Theory," *AI Alignment Forum*, Sept. 2020, [Online]. Available: https://www.alignmentforum.org/posts/basic-inframeasure-theory

[4] R. Nozick, "Newcomb's problem and two principles of choice," *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday*. Springer, pp. 114–146, 1969.