Imperfect World Models are Exploitable

Logan M. Bhamidipaty[1], **Esmeralda S. Whitammer**[1], **David Abel**[1], **Mykel J. Kochenderfer**[2], **Subramanian Ramamoorthy**[1]
l.m.bhamidipaty@ed.ac.uk

[1]**University of Edinburgh**
[2]**Stanford University**

Abstract

We propose a novel definition of model exploitation in reinforcement learning. Informally, a world model is exploitable if it implies that one policy should be strictly preferred over another while the environment's true transition model implies the reverse. We analogize our definition with a prior characterization of reward hacking but show that the associated proof of inevitability does not transfer to exploitation. To overcome this obstruction, we develop a general theory of reward hacking and model exploitation that proves that exploitation is essentially unavoidable on large policy sets and yields the corresponding claim for hacking as a special case. Unfortunately, we also find that the conditions that guarantee unhackability in finite policy sets have no counterpart that precludes exploitation. Consequently, we introduce a relaxed notion of exploitation and derive a safe horizon within which it can be avoided. Taken together, our results establish a formal bridge between reward hacking and model exploitation and elucidate the limits of safe planning in world models.

1 Introduction

Efficient real-world planning typically requires approximating complex dynamics (Simon, 1955; Javed and Sutton, 2024). In the best case, well-chosen approximations unravel otherwise intractable problems, as with admissible heuristics in search (Hart et al., 1968), model reduction in dynamical systems (Antoulas, 2005), and latent-space world models in sequential decision-making (Hafner et al., 2019). In the worst case, however, ill-chosen approximations can compromise safety, as with mimicry in evolutionary biology (Bates, 1862), arbitrage in financial markets (Ross, 1976), and speedrunning in video games (Scully-Blaker, 2014). In reinforcement learning (RL), this liability arises not only exogenously from adversaries but also endogenously from the optimizer itself: an agent trained to maximize total expected reward under an imperfect model[1] may discover behavior that performs well in simulation but poorly in reality (Jakobi et al., 1995; Tobin et al., 2017). This failure mode is called **model exploitation** (Ha and Schmidhuber, 2018; Janner et al., 2019), and a natural question is whether it can be avoided.

This question pertains to notable results in RL theory including the simulation lemma (Kearns and Singh, 2002) and the value equivalence principle (Grimm et al., 2020). Both implicitly use predictive performance to measure the quality of a model's approximation. The simulation lemma does so cardinally by bounding the error from optimizing under an imperfect model, while the value equivalence principle does so locally by characterizing when two models induce identical Bellman updates. Modern world models also emphasize predictive performance, typically learning latent-space models of transitions (Hafner et al., 2019; Balestriero and LeCun, 2025).

[1]We use "model" to mean the transition model of a Markov decision process (§2) or an agent's approximation to it, sometimes called a world model (Ha and Schmidhuber, 2018). We do not use it to mean the policy (as a generative model of actions) as is sometimes meant in other RL literature.

We argue that while this emphasis on performance is necessary for efficient planning, it is not sufficient for safe planning. Recent work on reward specification (Hadfield-Menell et al., 2017; Abel et al., 2021) views reward functions as observations of the designer's intent to communicate a goal, which may be an imperfect reflection of a true objective. Empirical transition models are similarly imperfect. In this setting, the natural notion of safety is ordinal rather than cardinal (*does the model preserve which policies are better than which?*) and global rather than local (*does it do so across the entire policy set, not just on the chosen classes?*). Concretely, a proxy model is *safe* in our sense if it does not invert the policy ordering induced by the true model. That is, whenever reality prefers a policy π over π', a safe approximation should not prefer π' over π. When such an inversion exists, we say the pair of transition models is *exploitable* (Definition 1).

Our formalization of model exploitation is analogous to the notion of reward hacking in Skalse et al. (2022). Unfortunately, their characterization of when hacking exists and is avoidable does not transfer to exploitation (§3.2). We overcome this by introducing a more general theory, from which four main contributions emerge.

1. We propose a new definition of model exploitation (Definition 1) that, to the best of our knowledge, is the first ordinal notion of safety for world models.
2. We develop a unified theory of reward hacking and model exploitation that characterizes exploitation on common policy sets (Corollaries 1 and 2), yields Theorem 1 of Skalse et al. (2022) as a special case (Corollary 3), and provides a shared geometric intuition for both phenomena (Lemmas 1 and 2).
3. We give a construction reducing any instance of model exploitation to one of reward hacking (Proposition 3), but show that a converse construction does not exist (Counterexample 1). We further prove that the conditions that guarantee unhackability are insufficient to guarantee unexploitability (Counterexample 2).
4. We introduce a relaxed notion of model exploitation, ε-exploitability (Definition 3), and use the tight form of the simulation lemma (Kearns and Singh, 2002) recently proved by Lobel and Parr (2024) to derive a safe horizon for discounted Markov decision processes within which ε-unexploitability is guaranteed (Theorem 2).

2 Preliminaries

We first recap the preliminaries for reinforcement learning (Sutton et al., 1998) and introduce our definition of model exploitation.

2.1 Reinforcement learning

We consider a **Markov decision process (MDP)** $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition model, $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution, $\mathcal{R}\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. As in Skalse et al. (2022), we assume finite $\mathcal{S}$ with all states reachable and finite $\mathcal{A}$ with $|\mathcal{A}| > 1$.

A **stationary policy** $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ maps each state to a distribution over actions. A **non-stationary policy** $\pi = (\pi_0, \pi_1, \pi_2, \dots)$ is a sequence of mappings $\pi_t\colon \mathcal{S} \to \Delta(\mathcal{A})$, allowing the action distribution to depend on the time step. Rolling out a policy produces a trajectory $\tau = (s_0, a_0, r_0, \dots)$ whose discounted return is $G(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t$. The **value** of a policy is its expected return $J(\pi) = \mathbb{E}_{\tau\sim\pi}[G(\tau)]$, or, equivalently, $J(\pi) = \langle \mathcal{R}, \mathcal{F}_{\mathcal{T}}^{\pi} \rangle$ where $\mathcal{F}_{\mathcal{T}}^{\pi}(s, a) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}(s_t = s, a_t = a)]$ are the **discounted visit counts** of π under $\mathcal{T}$. We make the standard assumption in classical RL and optimal control (Sutton et al., 1998; Bertsekas, 2012) that designers use value to compare policies and prefer high-value policies over low-value ones.

An **environment** (also sometimes called a controlled Markov process) is an MDP without a reward function $(\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, _, \gamma)$, and a **task** is an MDP without a transition model $(\mathcal{S}, \mathcal{A}, _, d_0, \mathcal{R}, \gamma)$. When considering multiple reward functions in a fixed environment, we write $J_{\mathcal{R}}(\pi) = \langle \mathcal{R}, \mathcal{F}^{\pi} \rangle$ (dropping the subscript on $\mathcal{F}^{\pi}$ when $\mathcal{T}$ is fixed). When considering multiple transition models in a

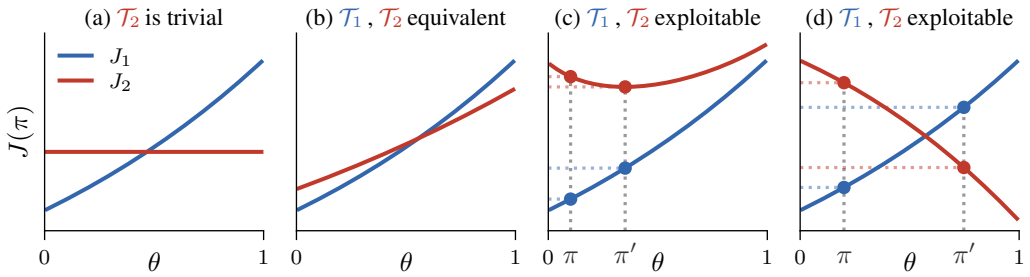

Figure 1: Taxonomy of transition model relationships in a 3-state MDP (Appendix F). Policies are parameterized by $\pi_\theta(a_0 \mid s) = \theta$ for all $s \in \mathcal{S}$. Each panel has the same $\mathcal{T}_1$ but a unique $\mathcal{T}_2$. **(a)** $\mathcal{T}_2$ is trivial, so J_2 is constant. **(b)** $\mathcal{T}_1$, $\mathcal{T}_2$ are equivalent (in policy ordering): both value functions increase in θ, preserving the policy ordering. **(c, d)** $\mathcal{T}_1$, $\mathcal{T}_2$ are exploitable: dotted lines show the exploiting pair (π, π') from Definition 1, with $J_1(\pi') > J_1(\pi)$ but $J_2(\pi) > J_2(\pi')$.

fixed task, we write $J_{\mathcal{T}}(\pi) = \langle \mathcal{R}, \mathcal{F}^{\pi}_{\mathcal{T}} \rangle$. We use J_i as shorthand for $J_{\mathcal{R}_i}$ or $J_{\mathcal{T}_i}$ depending on whether the index ranges over reward functions or transition models.

Since $\mathcal{S}$ and $\mathcal{A}$ are finite, the space of stationary policies can be represented as a product of $|\mathcal{S}|$ probability simplices $\Delta(\mathcal{A})^{|\mathcal{S}|}$. We write Π^+ for the interior of this space, consisting of all policies with $\pi(a \mid s) > 0$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Unless otherwise specified, openness is taken with respect to Π^+. We provide a reference list of notation and terminology in Appendix A.

2.2 Defining model exploitation and reward hacking

We define model exploitation as a binary relation between two transition models.

Definition 1 (Model Exploitation)**.** *Transition functions $\mathcal{T}$ and $\mathcal{T}'$ are **exploitable** relative to a policy set Π and a task $(\mathcal{S}, \mathcal{A}, _, d_0, \mathcal{R}, \gamma)$ if there exist $\pi, \pi' \in \Pi$ such that*

$$J_{\mathcal{T}}(\pi) > J_{\mathcal{T}}(\pi') \quad \textit{and} \quad J_{\mathcal{T}'}(\pi') > J_{\mathcal{T}'}(\pi),$$

*otherwise they are **unexploitable**.*

Intuitively, model exploitation says that optimizing under one transition model may look like a mistake under the other. This definition directly analogizes the notion of reward hacking in Skalse et al. (2022) with transition models in place of reward functions. We restate their definition below.

Definition 2 (Reward Hacking, Skalse et al. (2022))**.** *Reward functions $\mathcal{R}$ and $\mathcal{R}'$ are **hackable** relative to a policy set Π and an environment $(\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, _, \gamma)$ if there exist $\pi, \pi' \in \Pi$ such that*

$$J_{\mathcal{R}}(\pi) > J_{\mathcal{R}}(\pi') \quad \textit{and} \quad J_{\mathcal{R}'}(\pi') > J_{\mathcal{R}'}(\pi),$$

*otherwise they are **unhackable**.*

Both definitions are instances of a common primitive, which we call **value inversion**. We say two value functions J_1 and J_2 admit a **value inversion** on Π if there exist $\pi, \pi' \in \Pi$ such that $J_1(\pi) > J_1(\pi')$ and $J_2(\pi') > J_2(\pi)$. Thus, model exploitation is a value inversion between $J_{\mathcal{T}}$ and $J_{\mathcal{T}'}$, and reward hacking is a value inversion between $J_{\mathcal{R}}$ and $J_{\mathcal{R}'}$.

We further say J_1 and J_2 are **equivalent** on Π if they induce the same ordering on Π and that a value function J is **trivial** on Π if $J(\pi)$ is constant on Π (Figure 1). Informally, equivalence means that two value functions agree on every pairwise comparison between policies, while triviality means there are no meaningful comparisons to make. We also use trivial and equivalent to describe transition models and reward functions as shorthand for the corresponding properties of the value functions they induce. We note that triviality precludes value inversion and that value inversion is symmetric, irreflexive, and not necessarily transitive.

3 Results

To understand when model exploitation is avoidable, we ask a single question with increasing precision: *on which policy sets can we find non-trivial, non-equivalent, unexploitable transition model pairs?* We begin with the largest possible policy set and progressively tighten the scope.

3.1 Non-stationary policies

Consider first the set of all non-stationary policies Π_{NS}. If exploitation can be avoided here, it can be avoided on any policy set, since all policy sets are subsets of Π_{NS}. Unfortunately, this is false: we find that no interesting unexploitability exists.

Proposition 1 (No unexploitability on Π_{NS})**.** *On the set of all non-stationary policies, every non-trivial, non-equivalent pair of transition models is exploitable.*

We defer all formal proofs to Appendix B. The proof uses a similar construction to the one in Skalse et al. (2022), relying on the richness of Π_{NS} to find an exploiting policy pair, but says nothing about the set of stationary policies, where the question is considerably harder.

3.2 Distinguishing model exploitation from reward hacking

Having seen that interesting unexploitability is impossible in the most general policy set, a natural question is whether it can be avoided in smaller policy sets, such as deterministic or optimal policies. By Theorem 1 of Skalse et al. (2022), reward hacking is inevitable on any policy set that contains an open subset (under mild regularity conditions). Given that our proof of Proposition 1 follows a similar construction to the one in Skalse et al. (2022), it is reasonable to expect that we might derive analogous results using similar techniques. Unfortunately, this is not true, as we argue in this section.

3.2.1 Distinguishing transition models and reward functions

Transition models and reward functions have significantly different geometries that make it hard to reason about them in a unified manner. Two of these differences are important for our results. *First*, the policy value function $J(\pi) = \langle \mathcal{R}, \mathcal{F}^{\pi}_{\mathcal{T}} \rangle$ is linear in the reward function but nonlinear in the transition model's probability mass function. This means that the linear-algebraic and real-analytic machinery that Skalse et al. (2022) develop for their characterization of reward hacking generally does not transfer to the model exploitation setting, so analogous results (if they exist) must rely on different methods. *Second*, the only constraint imposed on the reward function is typically boundedness: rewards may not go to infinity. Transition models, on the other hand, are probability mass functions, which are nonnegative and sum to 1 at each state. This means reward functions live in the full Euclidean space $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, which is closed under negation, scaling, and addition, but transition models live on a compact product of simplices $\Delta(\mathcal{S})^{|\mathcal{S}||\mathcal{A}|}$, which does not have vector space structure (at least in the probability domain).

The primary consequence of these differences is that neither the techniques nor the results from reward hacking cleanly apply to model exploitation. In fact, as we show next, results that are true for hacking are sometimes false for exploitation.

3.2.2 Finite policy sets

Unlike reward hacking, model exploitation persists even in finite policy sets. Formally, Theorem 2 of Skalse et al. (2022) shows that, together with some mild regularity conditions, a non-trivial, non-equivalent unhackable reward pair always exists on any finite policy set.

Making the corresponding existence guarantee for unexploitable transition model pairs requires ruling out two degeneracies that have no analogue for reward hacking. When $\gamma = 0$, the value $J(\pi) = \sum_s d_0(s) \sum_a \pi(a|s)\, \mathcal{R}(s,a)$ is independent of $\mathcal{T}$, so every transition model pair is equivalent. The same obstruction applies when both $\mathcal{R}$ and every policy in Π are stateless (i.e., $\mathcal{R}(s,a) = \mathcal{R}(a)$ and

$\pi(a|s) = \pi(a)$ for all $s \in \mathcal{S}$), since then $J(\pi) = \frac{1}{1-\gamma} \sum_a \mathcal{R}(a)\, \pi(a)$. In both cases, no transition model pairs are non-equivalent.

In examining whether the remaining regularity conditions of Skalse et al. (2022) suffice once these degeneracies are excluded, we noticed that Theorem 2 requires a hypothesis that is not present in the original statement but implicitly used in the proof. (We view this as a minor oversight that does not affect the broader conclusions of their work, since the condition is degenerate and easily excluded.)

Proposition 2 (Non-collinearity is necessary for Theorem 2 of Skalse et al. (2022))**.** *On any policy set with collinear visit counts (i.e., all visit counts $\mathcal{F}^\pi$ for $\pi \in \Pi$ lie on a single line in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$), every non-trivial, non-equivalent pair of reward functions is hackable.*

We provide a corrected statement in Appendix E with the additional hypothesis that visit counts are not collinear. Unfortunately, even assuming all of the regularity conditions above together with those of Theorem 2, this is still insufficient to guarantee the existence of a non-trivial, non-equivalent unexploitable transition model pair on finite policy sets (see Counterexample 2 in Appendix B). We suspect that the finite-policy-set case admits no clean existence guarantee for unexploitability, because the nonlinear dependence of J on $\mathcal{T}$ makes the geometry of the induced value orderings fundamentally richer than in the reward hacking setting.

Question 1. *Find necessary and sufficient conditions on a task $(\mathcal{S}, \mathcal{A}, _, d_0, \mathcal{R}, \gamma)$ and finite policy set that guarantee the existence of a non-trivial, non-equivalent, unexploitable pair of transition models.*

3.3 Unifying model exploitation and reward hacking

We next return to our main question: *can exploitability be avoided by restricting the policy set?* As discussed in §3.2 above, Theorem 1 of Skalse et al. (2022) shows that the answer is no for reward hacking on any policy set containing an open subset, but the result does not directly apply to exploitation. Surprisingly, this does not mean that reward hacking has nothing to tell us about model exploitation. In this section, we first show that every instance of model exploitation can be reduced to an instance of reward hacking, then develop this insight into a unified theory of exploitation and hacking.

Proposition 3 (Exploitation implies hacking)**.** *For any exploitable pair $(\mathcal{T}, \mathcal{T}')$ relative to a policy set Π and a task $(\mathcal{S}, \mathcal{A}, _, d_0, \mathcal{R}, \gamma)$, there exists a reward function $\mathcal{R}'$ such that $(\mathcal{R}, \mathcal{R}')$ is hackable relative to Π and the environment $(\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, _, \gamma)$.*

While the result may be surprising, the proof is a short construction: set $\mathcal{R}'$ to be the difference in visit counts $\mathcal{F}_{\mathcal{T}}^{\pi'} - \mathcal{F}_{\mathcal{T}}^{\pi}$ for some exploiting policy pair $\pi, \pi' \in \Pi$. Interestingly, the converse does not hold.

Counterexample 1 (Hacking may not imply exploitation)**.** Consider any MDP with $|\mathcal{A}| \geq 2$ and $\gamma > 0$. Let $\mathcal{R}(s, a) = \mathbb{1}\{a = a_1\}$ and $\mathcal{R}'(s, a) = \mathbb{1}\{a = a_2\}$, and let π always select a_1 and π' always select a_2. Then $(\mathcal{R}, \mathcal{R}')$ is hackable, since $J_{\mathcal{R}}(\pi) = 1/(1-\gamma) > 0 = J_{\mathcal{R}}(\pi')$ and $J_{\mathcal{R}'}(\pi') = 1/(1-\gamma) > 0 = J_{\mathcal{R}'}(\pi)$, but no perturbation of the transition model can alter the relative ordering of π and π' under either reward function.

Proposition 3 is unassuming by itself, but it hints at something more fundamental: a unified theory of model exploitation and reward hacking may exist *despite* the anatomical differences between reward functions and transition models. Both model exploitation and reward hacking are inversions in the value ordering of a policy set. This suggests that the correct place to begin developing a unified treatment of exploitation and hacking is the value function. Unfortunately, the difficulties that make transition models unwieldy relative to reward functions (§3.2.1) also apply to the value function, so we still cannot apply the machinery of Skalse et al. (2022). The key insight to overcoming this obstacle is that the value function is rational in policy space.

Proposition 4 (Value function rationality)**.** *For any MDP with finite $\mathcal{S}$ and $\mathcal{A}$, the value function $J(\pi)$ is rational in π. In particular, J is real-analytic on Π^+.*

Rational functions are typically easier to reason about compared to nonlinear functions in general because they have a host of desirable properties (Rudin, 1974). The two properties that are useful for

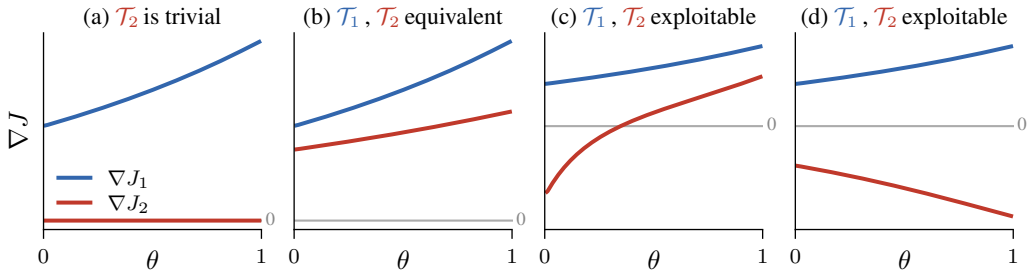

Figure 2: Gradients for the value curves in Figure 1. **(a)** ∇J_2 vanishes (trivial). **(b)** Both gradients positive: proportional gradients force equivalence (Lemma 2). **(c, d)** The gradients have opposite signs, yielding exploitation (Lemma 1). In (c) this occurs only near $\theta = 0$; in (d) it holds everywhere.

our purposes are analyticity and rigidity. The former provides access to an arsenal of powerful results from analysis and differential geometry, while the latter allows us to promote properties from open subsets to the supersets containing them.

More concretely, from these two properties, we can derive the following lemmas, which relate the gradients of value functions to the existence (or lack thereof) of inversions and give a simple geometric picture: at every point in policy space, the relationship between ∇J_1 and ∇J_2 determines whether the two value functions can disagree on which policy is better.[2]

Lemma 1 (Local inversion)**.** *Let Π be open. If there is a direction $v \in \mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$ along which J_1 increases and J_2 decreases at some $\pi \in \Pi$ (i.e., $\nabla_v J_2(\pi) < 0 < \nabla_v J_1(\pi)$), then J_1 and J_2 admit an inversion on Π.*

Lemma 2 (Global equivalence)**.** *Let $\tilde{\Pi} \subseteq \Pi$ be open. If J_1 and J_2 are non-trivial and have positively proportional gradients wherever both are nonzero on $\tilde{\Pi}$, then J_1 and J_2 are equivalent on Π.*

The intuition for Lemma 1 is as follows. If there is a direction in policy space that improves performance under one model but decreases it under the other, then a step in that direction leads to two policies that form an exploiting pair.

Lemma 2 has a similar intuition. If every direction that improves one value function also improves the other, then the two can never disagree on which of any two policies is better, so they must induce the same ordering. While the result is simple, the formal proof is surprisingly involved and relies heavily on value function rationality.

Together, these two lemmas cover every possible relationship between ∇J_1 and ∇J_2 (Figure 2). At every point in policy space, the gradients are either linearly independent, antiparallel, or positively proportional. The first two cases produce an inversion via Lemma 1; the third forces equivalence via Lemma 2. This gives our main characterization.

Theorem 1 (Value inversions)**.** *Let J_1 and J_2 be non-trivial, non-equivalent value functions. If the policy set contains an open subset, then it admits a value inversion for J_1 and J_2.*

The theorem immediately yields the desired characterization for model exploitation.

Corollary 1 (Imperfect world models are exploitable)**.** *On any policy set containing an open subset, every non-trivial, non-equivalent pair of transition models is exploitable.*

In practice, one might hope that Corollary 1 only applies to high-entropy or poorly performing policies that no reasonable learning algorithm would ever produce. Unfortunately, Theorem 1 still applies both to the set of ε-suboptimal policies (those π with $J(\pi) \geq \sup_{\pi'} J(\pi') - \varepsilon$ for $\varepsilon > 0$) and to the set of δ-deterministic policies (those π with $\max_a \pi(a \mid s) \geq \delta$ for $\delta < 1$ and every $s \in \mathcal{S}$).

Corollary 2 (Common exploitable policy sets)**.** *Every non-trivial, non-equivalent pair of transition models is exploitable on (i) the set of all stationary policies, (ii) the set of all ε-suboptimal policies ($\varepsilon > 0$), and (iii) the set of all δ-deterministic policies ($\delta < 1$).*

[2]From now on, we identify the tangent space to the policy space $\Delta(\mathcal{A})^{|\mathcal{S}|}$ at any π with $\mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$, although the identification is not canonical. We can thus write $\nabla_v J(\pi)$ for $v \in \mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$ an infinitesimal perturbation of π.

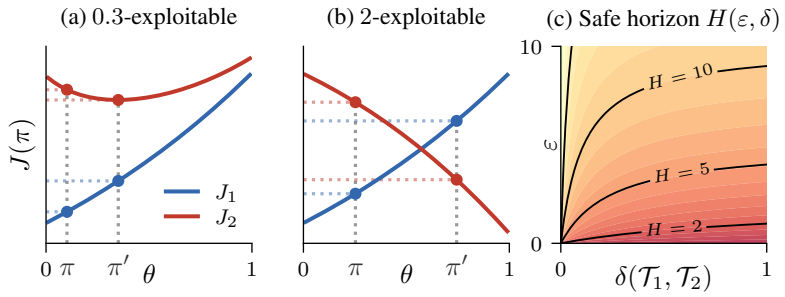

Figure 3: Examples of ε-exploitability and a contour plot for the safe horizon. **(a, b)** The exploitable transition model pairs from Figure 1(c, d), annotated with exploiting policies. The pair in (a) is 0.3-exploitable and the pair in (b) is 2-exploitable. **(c)** The safe horizon $H(\varepsilon, \delta)$ from Theorem 2. Each contour marks the largest effective horizon $1/(1-\gamma)$ under which ε-unexploitability is guaranteed for a given tolerance ε and maximum total variation distance $\delta(\mathcal{T}_1, \mathcal{T}_2)$.

Theorem 1 implies characterizations of both reward hacking and model exploitation. Skalse et al. (2022) arrive at the reward hacking result through an argument in the visit-count space, exploiting the linearity of J in the reward vector. The same conclusion follows *a fortiori* from Theorem 1, since the only property of J_1 and J_2 used in the proof is rationality in π, which holds regardless of whether the index ranges over reward functions or transition models.

Corollary 3 (Theorem 1 of Skalse et al. (2022))**.** *On any policy set containing an open subset, every non-trivial, non-equivalent pair of reward functions is hackable.*

3.4 Relaxation

We now introduce a relaxed notion of exploitability to pursue two goals. *First*, practitioners may not care about small-value inversions in real-world applications, so we seek a more graded measure of exploitability that distinguishes between negligible and catastrophic differences in value. *Second*, we want to answer our initial question about when exploitation is avoidable, even if only approximately. Thus far, we have seen that interesting unexploitability is impossible on large policy sets (Corollary 1) and resists characterization on finite ones (§3.2.2). A weaker definition, in contrast, may be more likely to admit positive existence conditions than our usual notion of exploitation.

Definition 3 (ε-Exploitation)**.** *Transition models $\mathcal{T}$ and $\mathcal{T}'$ are **ε-exploitable** relative to a policy set Π and a task $(\mathcal{S}, \mathcal{A}, _, d_0, \mathcal{R}, \gamma)$ if there exist $\pi, \pi' \in \Pi$ such that*

$$J_{\mathcal{T}}(\pi) - J_{\mathcal{T}}(\pi') > \varepsilon \quad \textit{and} \quad J_{\mathcal{T}'}(\pi') - J_{\mathcal{T}'}(\pi) > \varepsilon,$$

*otherwise they are **ε-unexploitable**.*

Intuitively, ε-unexploitability says that while small inversions may exist, $\mathcal{T}$ and $\mathcal{T}'$ may never confidently disagree about how to rank a policy pair. This definition immediately satisfies our first goal, because it generalizes exploitation (recovering Definition 1 when $\varepsilon = 0$) and is a continuous rather than binary measure of "exploitability" (i.e., via the minimum ε for which a pair of transition models remains ε-unexploitable).

To address our second goal and establish sufficient conditions for ε-unexploitability, we derive the following theorem.

Theorem 2 (Safe horizon)**.** *Let $\mathcal{R}\colon \mathcal{S} \times \mathcal{A} \to [0, 1]$ and let $\delta = \frac{1}{2} \max_{s,a} \|\mathcal{T}(\cdot|s,a) - \mathcal{T}'(\cdot|s,a)\|_1$ be the total variation distance between distinct $\mathcal{T}$ and $\mathcal{T}'$. For any $\varepsilon > 0$, define the safe horizon*

$$H(\varepsilon, \delta) = \frac{(1+\varepsilon) + \sqrt{(1-\varepsilon)^2 + 4\varepsilon/\delta}}{2}.$$

Every pair $(\mathcal{T}, \mathcal{T}')$ is ε-unexploitable on every policy set Π whenever $1/(1-\gamma) \leq H(\varepsilon, \delta)$. Furthermore, this bound is tight.

The safe horizon $H(\varepsilon, \delta)$ is the longest effective horizon $1/(1-\gamma)$ at which ε-unexploitability is guaranteed. Its dependence on both arguments aligns with standard intuitions in model-based RL. Increasing δ lowers H: a less accurate model exhausts its error budget sooner, so safe planning requires a shorter horizon. This is exactly in line with the main result by Jiang et al. (2015), who show that a shorter effective horizon should be used when data is scarce. In this sense, we corroborate and expand their findings through the lens of ε-exploitability.

Increasing ε raises H: a designer who tolerates larger inversions can plan further ahead. At the extremes, $H \to \infty$ as $\delta \to 0$ (a perfect model is never exploitable) and $H \to 1+\varepsilon$ as $\delta \to 1$ (a maximally inaccurate model is safe only near single-step planning). This formalizes the widely held intuition that model errors compound over the planning horizon (Ross et al., 2011; Talvitie, 2014), and provides a closed-form expression for exactly how much horizon a given level of model error can afford (Figure 3c).

The closed form of H is exact but unwieldy. A looser but more memorable sufficient condition follows from dropping lower-order terms.

Corollary 4 (Square-root heuristic)**.** *Let $\mathcal{R}\colon \mathcal{S} \times \mathcal{A} \to [0,1]$ and let $\delta = \frac{1}{2}\max_{s,a} \|\mathcal{T}(\cdot|s,a) - \mathcal{T}'(\cdot|s,a)\|_1 > 0$. For any $\varepsilon > 0$, the pair $(\mathcal{T}, \mathcal{T}')$ is ε-unexploitable whenever $1/(1-\gamma) < \sqrt{\varepsilon/\delta}$.*

Corollary 4 gives a quick sense of the magnitude of possible inversions: at $\gamma = 0.9$ the effective horizon is 10, so the worst-case inversion that cannot be ruled out scales as 100δ; at $\gamma = 0.99$ it scales as $10{,}000\delta$. Unlike the bound in Theorem 2, the approximate bound in Corollary 4 is not tight, so the actual inversion may be much smaller, but it provides an immediate sense of how model error and planning horizon interact without requiring further analysis. For the transition model pairs in Figure 1, we compare both bounds against the actual exploitation gaps in Appendix G.

4 History and etymology of model exploitation

The term "model exploitation" has no single origin in RL. Adjacent concepts, such as "model bias" and "compounding error," have appeared in the literature since at least the early 2010s (Deisenroth and Rasmussen, 2011; Ross et al., 2011), themselves inherited from control theory (Schneider, 1996; Kappen, 2005). As deep RL grew in popularity toward the end of that decade, practitioners turned to sample-efficient model-based RL (MBRL) to meet rising data demands. Learned models, however, are uncertain where data is scarce, and unlike in model-free methods, no trial-and-error mechanism exists to correct the resulting errors. Researchers began describing policies as "exploiting" these uncertain regions (Kurutach et al., 2018), though this usage of "exploit" also predates deep RL (cf. Schneider 1996). The set phrase "model exploitation" congealed only more recently, following Ha and Schmidhuber (2018) and Janner et al. (2019), and has since motivated a lineage of conservative offline MBRL methods (Yu et al., 2020; Kidambi et al., 2020). We provide a more complete related work in Appendix C.

5 Conclusion

There is increasing interest in approaches to decision-making that combine a forward predictive model with planning. Although there is substantial literature on the design and analysis of such model-based algorithms, the emergence of world models that make predictions directly from rich low-level sensor streams pushes the boundaries of these methods. This calls for a fresh look at the questions of model robustness and safety of the computed policies. In this paper, we formalize the concept of model exploitation and study the conditions under which policies trained on empirical models may be safe. We find that in sufficiently rich policy classes, exploitability is inevitable. However, under a relaxed definition, we can establish conditions for unexploitability. We suggest that this could guide the development of new approaches to safe policy synthesis, as well as parallel methods for quantifying the quality of synthesized policies with respect to such safety objectives.

We describe limitations and future work in Appendix D.

References

David Abel, Will Dabney, Anna Harutyunyan, Mark K Ho, Michael Littman, Doina Precup, and Satinder Singh. On the expressivity of Markov reward. *Neural Information Processing Systems (NeurIPS)*, 2021.

Athanasios C Antoulas. *Approximation of large-scale dynamical systems*. SIAM, 2005.

Dilip Arumugam, David Abel, Kavosh Asadi, Nakul Gopalan, Christopher Grimm, Jun Ki Lee, Lucas Lehnert, and Michael L Littman. Mitigating planner overfitting in model-based reinforcement learning. *arXiv preprint arXiv:1812.01129*, 2018.

Karl Johan Åström. Optimal control of Markov processes with incomplete state information i. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Henry Walter Bates. Contributions to an insect fauna of the amazon valley. Lepidoptera: Heliconidæ. *Transactions of the Linnean Society of London*, 23(3):495–566, 1862.

Dimitri Bertsekas. *Dynamic Programming and Optimal Control: Volume I*, volume 4. Athena scientific, 2012.

Marc Deisenroth and Carl E Rasmussen. PILCO: A model-based and data-efficient approach to policy search. *International Conference on Machine Learning (ICML)*, 2011.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. *Neural Information Processing Systems (NeurIPS)*, 2020.

Christopher Grimm, André Barreto, Greg Farquhar, David Silver, and Satinder Singh. Proper value equivalence. *Neural Information Processing Systems (NeurIPS)*, 2021.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Neural Information Processing Systems (NeurIPS)*, 2018.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. *Neural Information Processing Systems (NIPS)*, 2017.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *International Conference on Machine Learning (ICML)*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *International Conference on Learning Representations (ICLR)*, 2020.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. *International Conference on Learning Representations (ICLR)*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025a.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025b.

Lars Peter Hansen and Thomas J. Sargent. *Robustness*. Princeton University Press, 2008.

Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Nick Jakobi, Phil Husbands, and Inman Harvey. Noise and the reality gap: The use of simulation in evolutionary robotics. In *European Conference on Artificial Life*, pages 704–720. Springer, 1995.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Neural Information Processing Systems (NeurIPS)*, 2019.

Khurram Javed and Richard S Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. *Autonomous Agents and Multiagent Systems (AAMAS)*, 2015.

Hilbert J Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011–P11011, 2005.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. *Neural Information Processing Systems (NeurIPS)*, 2020.

Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. *International Conference on Learning Representations (ICLR)*, 2018.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. *International Conference on Learning Representations (ICLR)*, 2025.

Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. Goal misgeneralization in deep reinforcement learning. *International Conference on Machine Learning (ICML)*, 2021.

Sam Lobel and Ronald Parr. An optimal tightness bound for the simulation lemma. *Reinforcement Learning Conference (RLC)*, 2024.

Robert E Lucas, Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. North-Holland, 1976.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. *Artificial Intelligence and Statistics (AISTATS)*, 2011.

Stephen A Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3): 341–360, 1976.

Walter Rudin. *Real and complex analysis*. 1974.

Jeff Schneider. Exploiting model uncertainty estimates for safe dynamic control learning. *Neural Information Processing Systems (NIPS)*, 1996.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering Atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Rainforest Scully-Blaker. A practiced practice: Speedrunning through space with de certeau and virilio. *Game Studies*, 14(1):2016, 2014.

Herbert A Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, pages 99–118, 1955.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *Neural Information Processing Systems (NeurIPS)*, 2022.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. MIT Press Cambridge, 1998.

Erik Talvitie. Model regularization for stable sample rollouts. In *UAI*, pages 780–789, 2014.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *International Conference on Intelligent Robots and Systems (IROS)*, 2017.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. *Neural Information Processing Systems (NeurIPS)*, 2020.

A Notation

We first provide a table summarizing all relevant notation.

Notation	*Meaning*	*Definition*
$\mathcal{S}$	State space	
$\mathcal{A}$	Action space	
$\mathcal{T}$	Transition model	$\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
d_0	Initial state distribution	$d_0 \in \Delta(\mathcal{S})$
$\mathcal{R}$	Reward function	$\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
γ	Discount factor	$\gamma \in [0, 1)$
$\mathcal{M}$	Markov decision process	$(\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, \mathcal{R}, \gamma)$
$\mathcal{M} \setminus \mathcal{R}$	Environment (MDP without reward)	$(\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, _, \gamma)$
$\mathcal{M} \setminus \mathcal{T}$	Task (MDP without transitions)	$(\mathcal{S}, \mathcal{A}, _, d_0, \mathcal{R}, \gamma)$
π	Stationary policy	$\pi : \mathcal{S} \to \Delta(\mathcal{A})$
Π	Policy set	
Π^+	Interior of all stationary policies	$\{\pi \in \Delta(\mathcal{A})^{\lvert\mathcal{S}\rvert} : \pi(a\lvert s) > 0 \; \forall_{s\in\mathcal{S}, a\in\mathcal{A}}\}$
$\hat{\Pi}$	Finite policy set	
$\mathcal{F}^\pi_{\mathcal{T}}(s, a)$	Discounted visit counts	$\sum_{t=0}^{\infty} \gamma^t \mathbb{1}(s_t{=}s, a_t{=}a)$
$J(\pi)$	Policy value	$\langle \mathcal{R}, \mathcal{F}^\pi_{\mathcal{T}} \rangle$
$J_{\mathcal{T}}(\pi)$	Value under transition model $\mathcal{T}$	$\langle \mathcal{R}, \mathcal{F}^\pi_{\mathcal{T}} \rangle$
$J_{\mathcal{R}}(\pi)$	Value under reward function $\mathcal{R}$	$\langle \mathcal{R}, \mathcal{F}^\pi \rangle$
$J_i(\pi)$	Value under $\mathcal{R}_i$ or $\mathcal{T}_i$	(context determines index)
$V^\pi(s)$	State-value function	$\mathbb{E}_{\tau\sim\pi}[G(\tau) \mid s_0 = s]$
$r^\pi(s)$	Per-state expected reward	$\sum_a \pi(a\lvert s)\, \mathcal{R}(s, a)$
$(\mathcal{R}, \mathcal{R}')$ hackable	Reward hacking	$\exists_{\pi,\pi'\in\Pi}\; J_{\mathcal{R}}(\pi) > J_{\mathcal{R}}(\pi')$ and $J_{\mathcal{R}'}(\pi') > J_{\mathcal{R}'}(\pi)$
$(\mathcal{T}, \mathcal{T}')$ exploitable	Model exploitation	$\exists_{\pi,\pi'\in\Pi}\; J_{\mathcal{T}}(\pi) > J_{\mathcal{T}}(\pi')$ and $J_{\mathcal{T}'}(\pi') > J_{\mathcal{T}'}(\pi)$
Equivalent	Same policy ordering	$\forall_{\pi,\pi'\in\Pi}\; J_1(\pi) \geq J_1(\pi') \Leftrightarrow J_2(\pi) \geq J_2(\pi')$
Trivial	Constant value	$\forall_{\pi,\pi'\in\Pi}\; J(\pi) = J(\pi')$

Table 1: A summary of notation.

B Proofs

Proposition 1 (No unexploitability on Π_{NS})**.** *On the set of all non-stationary policies, every non-trivial, non-equivalent pair of transition models is exploitable.*

Proof of Proposition 1. The proof follows the same argument as the analogous result in Skalse et al. (2022), with transition models in place of reward functions. The key observation is that $J_i(\pi_\lambda)$ is affine in λ by linearity of expectation, regardless of whether i indexes reward functions or transition models. □

Proposition 2 (Non-collinearity is necessary for Theorem 2 of Skalse et al. (2022))**.** *On any policy set with collinear visit counts (i.e., all visit counts $\mathcal{F}^\pi$ for $\pi \in \Pi$ lie on a single line in $\mathbb{R}^{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}$), every non-trivial, non-equivalent pair of reward functions is hackable.*

Proof of Proposition 2. Collinearity means there exist $\pi_0, \pi_1 \in \Pi$ with $\mathcal{F}^{\pi_0} \neq \mathcal{F}^{\pi_1}$ such that $\mathcal{F}^{\pi} = \mathcal{F}^{\pi_0} + c_\pi(\mathcal{F}^{\pi_1} - \mathcal{F}^{\pi_0})$ for distinct scalars c_π. The value of π under any reward $\mathcal{R}$ is

$$J_{\mathcal{R}}(\pi) = J_{\mathcal{R}}(\pi_0) + c_\pi \langle \mathcal{R},\ \mathcal{F}^{\pi_1} - \mathcal{F}^{\pi_0} \rangle,$$

which is affine in c_π. Since the c_π are distinct, the ordering of Π under $\mathcal{R}$ is determined entirely by the sign of $\langle \mathcal{R}, \mathcal{F}^{\pi_1} - \mathcal{F}^{\pi_0} \rangle$. Given a non-trivial $\mathcal{R}$, any $\mathcal{R}'$ either has the same sign (same ordering, so equivalent), the opposite sign (reversed ordering, so hackable), or no sign ($\mathcal{R}'$ trivial on Π). A non-trivial, non-equivalent pair therefore cannot be unhackable on Π. □

Counterexample 2. Consider the task $(\mathcal{S}, \mathcal{A}, _, d_0, \mathcal{R}, \gamma)$ with $\mathcal{S} = \{s_0, s_1\}$, $\mathcal{A} = \{a_0, a_1\}$, $d_0 = (1/2,\ 1/2)$, $\mathcal{R}(s, a) = \mathbb{1}\{s = s_0,\ a = a_0\} + \mathbb{1}\{s = s_1,\ a = a_1\}$, and any $\gamma \in (0, 1)$. Let $\hat{\Pi} = \{\pi_0, \pi_1, \pi_2\}$ where π_0 plays a_0 at both states, π_1 plays a_0 at s_0 and a_1 at s_1, and π_2 plays a_1 at s_0 and a_0 at s_1.

This task satisfies $\gamma > 0$, $\mathcal{R}$ is non-trivial and state-action-dependent, and the visit counts of $\hat{\Pi}$ are non-collinear under any transition model: $\mathcal{F}^{\pi_0}$, $\mathcal{F}^{\pi_1}$, and $\mathcal{F}^{\pi_2}$ are supported on $\{(s_0, a_0), (s_1, a_0)\}$, $\{(s_0, a_0), (s_1, a_1)\}$, and $\{(s_0, a_1), (s_1, a_0)\}$ respectively, and no affine combination can reconcile the disjoint supports.

Under π_1, we see that $r^{\pi_1}(s) = 1$ for both states, so $J_{\mathcal{T}}(\pi_1) = 1/(1-\gamma)$ for every $\mathcal{T}$. Under π_2, we see that $r^{\pi_2}(s) = 0$ for both states, so $J_{\mathcal{T}}(\pi_2) = 0$ for every $\mathcal{T}$. Under π_0, $r^{\pi_0}(s_0) = 1$ and $r^{\pi_0}(s_1) = 0$, so $J_{\mathcal{T}}(\pi_0) = \mu_{\mathcal{T}}^{\pi_0}(s_0)$ where $\mu_{\mathcal{T}}^{\pi_0}(s)$ is the discounted visitation count of state s. Since $d_0(s_0) = 1/2$, the $t = 0$ contribution gives $\mu_{\mathcal{T}}^{\pi_0}(s_0) \geq 1/2 > 0$. Since $d_0(s_1) = 1/2$, we also have $\mu_{\mathcal{T}}^{\pi_0}(s_1) \geq 1/2$, so $\mu_{\mathcal{T}}^{\pi_0}(s_0) = 1/(1-\gamma) - \mu_{\mathcal{T}}^{\pi_0}(s_1) < 1/(1-\gamma)$. Therefore, for every $\mathcal{T}$,

$$\frac{1}{1-\gamma} = J_{\mathcal{T}}(\pi_1) > J_{\mathcal{T}}(\pi_0) > 0 = J_{\mathcal{T}}(\pi_2),$$

and every pair of transition models is equivalent on $\hat{\Pi}$.

Proposition 3 (Exploitation implies hacking)**.** *For any exploitable pair $(\mathcal{T}, \mathcal{T}')$ relative to a policy set Π and a task $(\mathcal{S}, \mathcal{A}, _, d_0, \mathcal{R}, \gamma)$, there exists a reward function $\mathcal{R}'$ such that $(\mathcal{R}, \mathcal{R}')$ is hackable relative to Π and the environment $(\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, _, \gamma)$.*

Proof of Proposition 3. Since $(\mathcal{T}, \mathcal{T}')$ is exploitable, there exist $\pi, \pi' \in \Pi$ with $J_{\mathcal{T}}(\pi) > J_{\mathcal{T}}(\pi')$ and $J_{\mathcal{T}'}(\pi') > J_{\mathcal{T}'}(\pi)$. Working in the environment $(\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, _, \gamma)$, the original reward $\mathcal{R}$ satisfies $J_{\mathcal{R}}(\pi) = J_{\mathcal{T}}(\pi) > J_{\mathcal{T}}(\pi') = J_{\mathcal{R}}(\pi')$. Define $\mathcal{R}' = \mathcal{F}_{\mathcal{T}}^{\pi'} - \mathcal{F}_{\mathcal{T}}^{\pi}$. Since $J_{\mathcal{T}}(\pi) \neq J_{\mathcal{T}}(\pi')$, the visit counts $\mathcal{F}_{\mathcal{T}}^{\pi}$ and $\mathcal{F}_{\mathcal{T}}^{\pi'}$ are distinct, so $\mathcal{R}' \neq 0$. Then

$$J_{\mathcal{R}'}(\pi') - J_{\mathcal{R}'}(\pi) = \langle \mathcal{R}', \mathcal{F}_{\mathcal{T}}^{\pi'} - \mathcal{F}_{\mathcal{T}}^{\pi} \rangle = \|\mathcal{F}_{\mathcal{T}}^{\pi'} - \mathcal{F}_{\mathcal{T}}^{\pi}\|^2 > 0.$$

Therefore $J_{\mathcal{R}}(\pi) > J_{\mathcal{R}}(\pi')$ and $J_{\mathcal{R}'}(\pi') > J_{\mathcal{R}'}(\pi)$, so $(\mathcal{R}, \mathcal{R}')$ is hackable. □

Proposition 4 (Value function rationality)**.** *For any MDP with finite $\mathcal{S}$ and $\mathcal{A}$, the value function $J(\pi)$ is rational in π. In particular, J is real-analytic on Π^+.*

Proof of Proposition 4. Write $J(\pi) = \sum_s d_0(s)\, V^\pi(s)$; it suffices to show each $V^\pi(s)$ is rational in π. Let $\mathcal{R}^\pi(s) = \sum_a \pi(a|s)\, \mathcal{R}(s, a)$, $P^\pi(s'|s) = \sum_a \pi(a|s)\, \mathcal{T}(s'|s, a)$, and $A = I - \gamma P^\pi$. The Bellman equation in matrix form is $AV^\pi = \mathcal{R}^\pi$. Since P^π is a probability matrix and $\gamma \in [0, 1)$, A is invertible and $V^\pi = A^{-1}\mathcal{R}^\pi$. By Cramer's rule, V^π can then be expressed as a ratio of determinants and is thus rational as required. □

Lemma 1 (Local inversion)**.** *Let Π be open. If there is a direction $v \in \mathbb{R}^{|\mathcal{S}|(|\mathcal{A}|-1)}$ along which J_1 increases and J_2 decreases at some $\pi \in \Pi$ (i.e., $\nabla_v J_2(\pi) < 0 < \nabla_v J_1(\pi)$), then J_1 and J_2 admit an inversion on Π.*

Proof of Lemma 1. Define $\pi^{\pm} = \pi \pm \varepsilon v$ for $\varepsilon > 0$. Since each J_i is analytic (Proposition 4), first-order Taylor expansion gives $J_i(\pi^{\pm}) = J_i(\pi) \pm \varepsilon\langle \nabla J_i(\pi), v\rangle + O(\varepsilon^2)$ and subtracting yields

$$J_1(\pi^+) - J_1(\pi^-) = 2\varepsilon\langle \nabla J_1(\pi), v\rangle + O(\varepsilon^2)$$
$$J_2(\pi^-) - J_2(\pi^+) = -2\varepsilon\langle \nabla J_2(\pi), v\rangle + O(\varepsilon^2)$$

Since $\langle \nabla J_1(\pi), v\rangle > 0$ and $\langle \nabla J_2(\pi), v\rangle < 0$, we can make ε small enough so that both differences are positive and $\pi^{\pm} \in \Pi$ (openness). Thus, $\mathcal{T}_1$ and $\mathcal{T}_2$ are exploitable. □

Proposition 5 (Connected level sets)**.** *For every transition model $\mathcal{T}$ and $c \in \mathbb{R}$, the level set $\{J_{\mathcal{T}} = c\} \cap \Pi^+$ is connected.*

Proof of Proposition 5. Since the map $\pi \mapsto \mathcal{F}^{\pi}$ is a homeomorphism on Π^+ (Skalse et al., 2022, Lemma 1), it suffices to show connectedness in visit-count space. The set $\mathcal{F}(\Pi^+)$ is the interior of a polyhedron (defined by non-negativity and Bellman flow constraints), hence convex. Since $J_{\mathcal{T}}(f) = \langle \mathcal{R}, f\rangle$ is linear, $\{\langle \mathcal{R}, f\rangle = c\} \cap \mathcal{F}(\Pi^+)$ is a hyperplane intersected with a convex set, which is convex and hence connected. □

Lemma 2 (Global equivalence)**.** *Let $\tilde{\Pi} \subseteq \Pi$ be open. If J_1 and J_2 are non-trivial and have positively proportional gradients wherever both are nonzero on $\tilde{\Pi}$, then J_1 and J_2 are equivalent on Π.*

Proof of Lemma 2. We show that positive proportionality of gradients on $\tilde{\Pi}$ forces equivalence on all of Π. The idea is to first extend proportionality from $\tilde{\Pi}$ to all of Π^+ using analyticity and connectedness, then construct a strictly increasing reparametrization $J_2 = \varphi \circ J_1$ on Π^+, and finally extend this to Π by density.

We begin by extending the proportionality to Π^+. We want to establish that $\nabla J_2 = \lambda \nabla J_1$ with $\lambda > 0$ on all of Π^+. This requires showing that on Π^+ (i) the gradients are linearly dependent, (ii) the gradients are nonzero, and (iii) λ is positive.

For (i), the Gram determinant of ∇J_1 and ∇J_2 is real analytic on the domain Π^+ (Proposition 4) and vanishes on the open subset $\tilde{\Pi}$, so it vanishes identically on Π^+. The gradients are therefore linearly dependent everywhere on Π^+.

For (ii), write $f_i \colon \pi \mapsto \mathcal{F}^{\pi}_{\mathcal{T}_i}$. The chain rule gives $\nabla J_i(\pi) = Df_i(\pi)^{\top}\mathcal{R}_i$. The map f_i is a homeomorphism from Π^+ onto its image (Skalse et al., 2022, Lemma 1). Both f_i and its inverse $\mathcal{F} \mapsto \pi(a|s) = \mathcal{F}(s,a)/\sum_{a'} \mathcal{F}(s,a')$ are smooth (the former by rationality, the latter since all visitations are positive on Π^+), so f_i is in fact a diffeomorphism onto its image (in particular, a smooth immersion) and $Df_i(\pi)$ has full column rank. Non-triviality of J_i ensures $\mathcal{R}$ is not orthogonal to the column space of $Df_i(\pi)$, since otherwise J_i would be constant. Therefore $\nabla J_i(\pi) = Df_i(\pi)^{\top}\mathcal{R} \neq 0$ on Π^+.

For (iii), since both gradients are nonzero we may write $\nabla J_2 = \lambda \nabla J_1$ where $\lambda(\pi) = \langle \nabla J_2, \nabla J_1\rangle / \|\nabla J_1\|^2$ is continuous and nonzero on Π^+. Since Π^+ is connected, λ has constant sign. Since $\lambda > 0$ on $\tilde{\Pi}$ by hypothesis, $\lambda > 0$ on Π^+.

Having extended positive proportionality of the gradients to Π^+, we now construct a strictly increasing φ with $J_2 = \varphi \circ J_1$ on Π^+. Such a φ exists if and only if $J_1(\pi) = J_1(\pi')$ implies $J_2(\pi) = J_2(\pi')$ for all $\pi, \pi' \in \Pi^+$. Pick any such pair π, π'. Both policies lie in the same level set of J_1 in Π^+, and the level set is connected (Proposition 5), so there is a path $\alpha : [0,1] \to \Pi^+$ in the level set with $\alpha(0) = \pi$ and $\alpha(1) = \pi'$. Since $J_1(\alpha(t))$ is constant, differentiating gives $\langle \nabla J_1(\alpha(t)), \frac{d}{dt}\alpha(t)\rangle = 0$. Since $\nabla J_2 = \lambda \nabla J_1$, we also have $\langle \nabla J_2(\alpha(t)), \frac{d}{dt}\alpha(t)\rangle = 0$, so $J_2(\alpha(t))$ is constant on $[0,1]$, giving $J_2(\pi) = J_2(\alpha(0)) = J_2(\alpha(1)) = J_2(\pi')$.

It remains to show that such a φ is strictly increasing on the image of J_1. Fix $c \in J_1(\Pi^+)$ and pick π_0 with $J_1(\pi_0) = c$. Since $\nabla J_1(\pi_0) \neq 0$, we can choose a direction v with $\langle \nabla J_1(\pi_0), v\rangle \neq 0$ and define $h(x) = J_1(\pi_0 + xv)$. Then $h(0) = c$ and $\frac{d}{dx}h(0) \neq 0$, so by the inverse function theorem h is

locally invertible. Setting $\beta(t) = \pi_0 + h^{-1}(t)v$ gives a smooth curve with $J_1(\beta(t)) = h(h^{-1}(t)) = t$ near c. Then $\varphi(t) = J_2(\beta(t))$, so

$$\frac{d}{dt}\varphi(t) = \langle \nabla J_2(\beta(t)), \tfrac{d}{dt}\beta(t)\rangle = \lambda(\beta(t))\langle \nabla J_1(\beta(t)), \tfrac{d}{dt}\beta(t)\rangle = \lambda(\beta(t)),$$

where the last equality uses $\frac{d}{dt}J_1(\beta(t)) = 1$. Since $\lambda > 0$, φ is locally strictly increasing at every point of $J_1(\Pi^+)$, hence strictly increasing on $J_1(\Pi^+)$.

Finally, we use the density of Π^+ in the space of all stationary policies to extend to Π. Define $\bar{\varphi}(u) = \lim_{n\to\infty} \varphi(u_n)$ for any sequence $(u_n) \subseteq J_1(\Pi^+)$ with $u_n \to u$; this is well-defined because φ is bounded and monotone. By continuity of J_2, we have $\bar{\varphi}(J_1(\pi)) = \lim_{n\to\infty} \varphi(J_1(\pi_n)) = \lim_{n\to\infty} J_2(\pi_n) = J_2(\pi)$ for any $\pi \in \Pi$ and any sequence $(\pi_n) \subseteq \Pi^+$ with $\pi_n \to \pi$. The extension $\bar{\varphi}$ is strictly increasing: given $a < d$ in $J_1(\Pi)$, density of $J_1(\Pi^+)$ provides $b, c \in J_1(\Pi^+)$ with $a < b < c < d$, giving $\bar{\varphi}(a) \le \varphi(b) < \varphi(c) \le \bar{\varphi}(d)$. So for any $\pi, \pi' \in \Pi$, $J_1(\pi) > J_1(\pi')$ implies $J_2(\pi) = \bar{\varphi}(J_1(\pi)) > \bar{\varphi}(J_1(\pi')) = J_2(\pi')$, and J_1 and J_2 induce the same ordering on Π, contradicting non-equivalence. □

Theorem 1 (Value inversions)**.** *Let J_1 and J_2 be non-trivial, non-equivalent value functions. If the policy set contains an open subset, then it admits a value inversion for J_1 and J_2.*

Proof of Theorem 1. Let $\tilde{\Pi} \subseteq \Pi$ be open, and let $\mathcal{T}_1$, $\mathcal{T}_2$ be non-trivial and non-equivalent on Π. Since J_1 and J_2 are rational (Proposition 4), their gradients ∇J_1 and ∇J_2 exist. If these gradients are linearly independent at some $\pi_0 \in \tilde{\Pi}$, we can solve $\langle \nabla J_1(\pi_0), v\rangle = 1$ and $\langle \nabla J_2(\pi_0), v\rangle = -1$ (a consistent linear system in $|\mathcal{S}|(|\mathcal{A}| - 1) \ge 2$ unknowns), and Lemma 1 gives exploitation. If instead they are everywhere linearly dependent on $\tilde{\Pi}$, we may write $\nabla J_2 = \lambda \nabla J_1$ for a scalar-valued function λ wherever both gradients are nonzero (such points exist since each ∇J_i is analytic and not identically zero by non-triviality, so it is nonzero on a subset open and dense in $\tilde{\Pi}$). If $\lambda(\pi_0) < 0$ at some such point, setting $v = \nabla J_1(\pi_0)$ gives directional derivatives of opposite sign, and Lemma 1 again gives exploitation. The remaining possibility is that $\lambda > 0$ wherever both gradients are nonzero on $\tilde{\Pi}$, but then Lemma 2 forces $\mathcal{T}_1$ and $\mathcal{T}_2$ to be equivalent on Π, contradicting our assumption. □

Theorem 2 (Safe horizon)**.** *Let $\mathcal{R}\colon \mathcal{S} \times \mathcal{A} \to [0, 1]$ and let $\delta = \frac{1}{2}\max_{s,a} \|\mathcal{T}(\cdot|s, a) - \mathcal{T}'(\cdot|s, a)\|_1$ be the total variation distance between distinct $\mathcal{T}$ and $\mathcal{T}'$. For any $\varepsilon > 0$, define the safe horizon*

$$H(\varepsilon, \delta) = \frac{(1 + \varepsilon) + \sqrt{(1 - \varepsilon)^2 + 4\varepsilon/\delta}}{2}.$$

Every pair $(\mathcal{T}, \mathcal{T}')$ is ε-unexploitable on every policy set Π whenever $1/(1 - \gamma) \le H(\varepsilon, \delta)$. Furthermore, this bound is tight.

Proof of Theorem 2. Write $h = 1/(1 - \gamma)$. The tight simulation lemma (Lobel and Parr, 2024) with $\varepsilon_{\mathcal{R}} = 0$ gives $|J_1(\pi) - J_2(\pi)| \le B$ for every π, where $B = h - \frac{h}{1-(1-h)\delta}$.

Suppose $(\mathcal{T}_1, \mathcal{T}_2)$ is ε-exploitable via π, π'. Add the two defining inequalities:

$$\underbrace{(J_1(\pi) - J_2(\pi))}_{\le B} + \underbrace{(J_2(\pi') - J_1(\pi'))}_{\le B} \;>\; 2\varepsilon,$$

so $B > \varepsilon$. Contrapositively, $B \le \varepsilon$ implies ε-unexploitability.

It is elementary to show that for $\delta, \varepsilon > 0$ there is a unique $h = H(\varepsilon, \delta) > 1$ at which $B = \varepsilon$, and $h < H(\varepsilon, \delta)$ implies $B < \varepsilon$. Solving $B = \varepsilon$ reduces to a quadratic whose positive root is the $H(\varepsilon, \delta)$ in the theorem statement.

For tightness, take $\mathcal{S} = \{0, 1\}$, $\mathcal{A} = \{0, 1\}$, $d_0(0) = 1$, $\mathcal{R}(s, a) = \mathbb{1}\{s = 0\}$ and let state 1 be absorbing. Under $\mathcal{T}_1$, action 0 self-loops at state 0 and action 1 transitions to state 1 with probability δ. Under $\mathcal{T}_2$ the actions swap, so action 1 self-loops at state 0 and action 0 transitions to state 1 with probability δ. Then $J_1(\pi_0) = J_2(\pi_1) = h$ and $J_1(\pi_1) = J_2(\pi_0) = h/(1 + (h-1)\delta)$, so both gaps equal B exactly, and the pair is ε-exploitable whenever $h > H$. □

Corollary 4 (Square-root heuristic)**.** *Let* $\mathcal{R}\colon \mathcal{S} \times \mathcal{A} \to [0,1]$ *and let* $\delta = \frac{1}{2}\max_{s,a} \|\mathcal{T}(\cdot|s,a) - \mathcal{T}'(\cdot|s,a)\|_1 > 0$*. For any* $\varepsilon > 0$*, the pair* $(\mathcal{T}, \mathcal{T}')$ *is* ε*-unexploitable whenever* $1/(1-\gamma) < \sqrt{\varepsilon/\delta}$*.*

Proof of Corollary 4. It suffices to show $\sqrt{\varepsilon/\delta} \leq H(\varepsilon, \delta)$, since then $1/(1-\gamma) < \sqrt{\varepsilon/\delta}$ implies $1/(1-\gamma) < H(\varepsilon, \delta)$ and Theorem 2 gives ε-unexploitability. Since $(1-\varepsilon)^2 \geq 0$,

$$\sqrt{(1-\varepsilon)^2 + \frac{4\varepsilon}{\delta}} \;\geq\; 2\sqrt{\frac{\varepsilon}{\delta}},$$

so $H(\varepsilon, \delta) \geq \frac{1+\varepsilon}{2} + \sqrt{\varepsilon/\delta} > \sqrt{\varepsilon/\delta}$. □

C Extended related work

Other adjacent phenomena. Model exploitation overlaps with several independent concepts. Planner overfitting (Arumugam et al., 2018; Jiang et al., 2015) refers to a planner discovering spurious shortcuts in a learned model and develops regularization-based remedies. Jiang et al. (2015) were the first to formalize and isolate this phenomenon by highlighting the role that the effective horizon plays in regularizing a learned transition model when data is scarce. The sim-to-real gap (Jakobi et al., 1995; Tobin et al., 2017) describes the same failure when transferring controllers from simulated to physical environments, typically addressed through domain randomization or system identification. Objective robustness (Langosco et al., 2021) studies agents that achieve high reward in training environments but pursue unintended objectives at deployment, which can be viewed as exploitation of a training-time model of the deployment distribution. Lastly, as discussed, the value equivalence principle (Grimm et al., 2020; 2021) provides a framing of when two models can be well-understood as equivalent in terms of *value preservation* under the Bellman operator. This principle is closely tied to our notion of equivalent transition models (Figure 1).

MDP extensions. Several lines of work extend the MDP formalism to defend against imperfect transition models, typically by modifying the optimization problem rather than studying the misspecification itself. Robust MDPs (Iyengar, 2005; Nilim and El Ghaoui, 2005) represent transition uncertainty through uncertainty sets and optimize worst-case value, a cardinal safety criterion rather than the ordinal one we propose. MOReL (Kidambi et al., 2020) constructs a pessimistic MDP by introducing absorbing states at high-uncertainty transitions, and provides strong suboptimality guarantees for the resulting policy, but does not formalize the notion of exploitation it defends against. Both modify the MDP to cope with model uncertainty; our work instead characterizes when that uncertainty inverts the policy ordering.

World models. A growing family of methods trains policies entirely inside learned dynamics models, from the Dreamer line (Hafner et al., 2020; 2021; 2025a;b) to MuZero (Schrittwieser et al., 2020) to recent JEPA-style predictive representations (Balestriero and LeCun, 2025). These methods measure model quality through predictive accuracy or downstream task performance, both of which are cardinal. Our results show that low prediction error does not preclude ordinal failures: a model can be accurate yet still invert the policy ordering induced by the true dynamics (Theorem 2).

Outside machine learning. Optimizing against an imperfect model is a recurring failure mode across disciplines, so there is an extensive literature that exists outside machine learning. For example, in macroeconomics, the Lucas critique (Lucas, 1976) warns that policy rules derived from an estimated model can fail when agents adapt to the policy itself, and robust control (Hansen and Sargent, 2008) responds by optimizing against worst-case model perturbations. In sim-to-real robotics, controllers optimized in simulation routinely exploit discrepancies with real-world physics (Jakobi et al., 1995; Tobin et al., 2017). Our formalization of exploitation as a value inversion applies in principle to any setting where a proxy model guides optimization over a structured decision space.

D Limitations and future work

We identify three main limitations of our work. *First*, our results, like those in Skalse et al. (2022), concern MDPs with finite state and action spaces. We defer extensions to continuous and partially observable MDPs (POMDPs) to future work, noting that solving the former may yield the latter (since every POMDP can be written as a continuous belief-state MDP (Åström, 1965)). *Second*, also like in Skalse et al. (2022), our notion of exploitation is binary and restrictive. In practice, it is unlikely that small, rare value inversions will pose meaningful safety concerns. Future work could refine our notion of exploitation by introducing a more granular measure of exploitability (e.g., deriving bounds using correlation metrics (Laidlaw et al., 2025) or expanding upon ε-exploitation). *Third*, there are, as of yet, no known closed-form necessary and sufficient conditions that guarantee unexploitability in finite policy sets. Although we question whether such simple conditions exist (Question 1), there is still area to explore. Researchers interested in pursuing this direction may, for instance, find it productive to begin work on special finite policy sets and specific MDPs (e.g., characterize exploitation on deterministic policies in linear MDPs).

E Relevant reward hacking theorems

For convenience, we restate the two main characterization results from Skalse et al. (2022) in our notation. Note that Corollary 3 and Theorem S1 are contrapositives of the same fact: Theorem S1 says that any unhackable, non-trivial pair must be equivalent, while Corollary 3 says that any non-trivial, non-equivalent pair must be hackable. We use the latter phrasing to parallel our exploitability results.

Theorem S1 (Theorem 1 of Skalse et al. (2022))**.** *In any environment* $(\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, _, \gamma)$*, if* $\hat{\Pi}$ *contains an open set, then any pair of reward functions that are unhackable and non-trivial on* $\hat{\Pi}$ *are equivalent on* $\hat{\Pi}$*.*

As shown in Proposition 2, the original statement of the following result requires an additional hypothesis not present in the original: that the visit counts of $\hat{\Pi}$ are not collinear. The correction is minor and the strengthened condition is easily satisfied in practice; the core insights of Skalse et al. (2022) are unaffected.

Theorem S2 (Theorem 2 of Skalse et al. (2022), corrected)**.** *For any environment* $(\mathcal{S}, \mathcal{A}, \mathcal{T}, d_0, _, \gamma)$*, any finite set of policies* $\hat{\Pi}$ *whose visit counts are not collinear, and any reward function* $\mathcal{R}_1$*, there is a non-trivial reward function* $\mathcal{R}_2$ *such that* $\mathcal{R}_1$ *and* $\mathcal{R}_2$ *are unhackable but not equivalent.*

Proof. The proof is identical to that of Skalse et al. (2022). The only step requiring the strengthened hypothesis is ensuring that there is a path from $\vec{\mathcal{R}}_1$ to $-\vec{\mathcal{R}}_1$ along which the first reward function where some inequality becomes an equality is not trivial. With collinear visit counts, every pairwise value difference is a scalar multiple of a single inner product $\langle \mathcal{R}, \mathcal{F}^{\pi_1} - \mathcal{F}^{\pi_0} \rangle$, so zeroing any one difference zeros them all. Non-collinearity guarantees three policies π_a, π_b, π_c with $\mathcal{F}^{\pi_b} - \mathcal{F}^{\pi_a}$ and $\mathcal{F}^{\pi_c} - \mathcal{F}^{\pi_a}$ linearly independent. The set of rewards trivial on $\hat{\Pi}$ must be orthogonal to both, giving a subspace of codimension at least 2, which cannot disconnect $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. A path from $\vec{\mathcal{R}}_1$ to $-\vec{\mathcal{R}}_1$ therefore exists that avoids trivial rewards entirely, and the remainder of the argument proceeds unchanged. □

F Transition models for Figures 1 to 3

The MDP used in Figures 1 to 3 has $\mathcal{S} = \{s_0, s_1, s_2\}$, $\mathcal{A} = \{a_0, a_1\}$, $\mathcal{R}(s, a) = \mathbb{1}\{s = s_0\}$, $\gamma = 0.9$, and $d_0 = (1/3,\ 1/3,\ 1/3)$. Policies are parameterized by $\pi_\theta(a_0 \mid s) = \theta$ for all $s \in \mathcal{S}$. Each transition model is specified by two stochastic matrices P_{a_0} and P_{a_1}, where entry (i, j) gives $\mathcal{T}(s_j \mid s_i, a)$.

The shared transition model $\mathcal{T}_1$ is

$$P_{a_0}^{(1)} = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.3 & 0.3 \end{pmatrix}, \qquad P_{a_1}^{(1)} = \begin{pmatrix} 0.1 & 0.3 & 0.6 \\ 0.1 & 0.2 & 0.7 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}.$$

Under $\mathcal{T}_1$, action a_0 concentrates transitions toward s_0 (the rewarding state), while action a_1 pushes toward s_2, making J_1 monotone increasing in θ.

Panel (a): $\mathcal{T}_2$ trivial. Both actions produce identical transitions, so J_2 is constant in θ:

$$P_{a_0}^{(2a)} = P_{a_1}^{(2a)} = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}.$$

Panel (b): $\mathcal{T}_1$, $\mathcal{T}_2$ equivalent. We construct $\mathcal{T}_2$ by blending $\mathcal{T}_1$ toward the uniform kernel with mixing weight $\alpha = 0.3$, i.e., $\mathcal{T}_2(s' \mid s, a) = 0.7 \cdot \mathcal{T}_1(s' \mid s, a) + 0.3/|\mathcal{S}|$:

$$P_{a_0}^{(2b)} = \begin{pmatrix} 0.59 & 0.24 & 0.17 \\ 0.45 & 0.31 & 0.24 \\ 0.38 & 0.31 & 0.31 \end{pmatrix}, \qquad P_{a_1}^{(2b)} = \begin{pmatrix} 0.17 & 0.31 & 0.52 \\ 0.17 & 0.24 & 0.59 \\ 0.17 & 0.17 & 0.66 \end{pmatrix}.$$

This preserves the qualitative structure of $\mathcal{T}_1$ (and thus the policy ordering) while attenuating the transition probabilities.

Panels (c) and (d): $\mathcal{T}_1$, $\mathcal{T}_2$ exploitable. The transition models $\mathcal{T}_2^{(c)}$ and $\mathcal{T}_2^{(d)}$ were found by computational search over random transition matrices drawn from a Dirichlet distribution. Full numerical values and the search code are available in the supplementary material.

G Bound comparison

Panel	Relationship	δ	Actual gap	Theorem 2 bound (tight)	Corollary 4 bound (loose)
(a)	Trivial	0.40	0	7.83	40
(b)	Equivalent	0.14	0	5.58	14
(c)	Exploitable	0.81	0.66	8.79	81
(d)	Exploitable	0.64	4.35	8.52	64

Table 2: Exploitation gaps for the transition model pairs in Figure 1 at $\gamma = 0.9$. The actual gap is the largest ε for which the pair is ε-exploitable. The Theorem 2 and Corollary 4 bounds give the unique ε above which each result guarantees ε-unexploitability. Both bounds are conservative.

Table 2 compares both bounds against the actual exploitation gaps for the transition model pairs in Figure 1. Both thresholds are conservative in every case and overestimate the actual gap. This conservatism is unavoidable in general: we prove Theorem 2 is tight.