How to Lose Inherent Counterfactuality in Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Learning in high-dimensional MDPs with complex state dynamics became possible with the progress achieved in reinforcement learning research. At the same time, deep neural policies have been observed to be highly unstable with respect to the minor variations in their state space, causing volatile and unpredictable behaviour. To alleviate these volatilities, a line of work suggested techniques to cope with this problem via explicitly regularizing the temporal difference loss to ensure local ϵ-invariance in the state space. In this paper, we provide theoretical foundations on the impact of ϵ-local invariance training on the deep neural policy manifolds. Our comprehensive theoretical and experimental analysis reveals that standard reinforcement learning inherently learns counterfactual values while recent training techniques that focus on explicitly enforcing ϵ-local invariance cause policies to lose counterfactuality, and further result in learning misaligned and inconsistent values. In connection to this analysis, we further highlight that this line of training methods break the core intuition and the true biological inspiration of reinforcement learning, and introduce an intrinsic gap between how natural intelligence understands and interacts with an environment in contrast to AI agents trained via ϵ-local invariance methods. The misalignment, inaccuracy and the loss of counterfactuality revealed in our paper further demonstrate the need to rethink the approach in establishing truly reliable and generalizable reinforcement learning policies.

1 Introduction

Inspired by the learning dynamics and cognitive abilities of natural intelligence (Watkins, 1989; Schmidhuber, 1999; Kehoe et al., 1987; Romo & Schultz, 1990; Montague et al., 1996; Schultz et al., 1993; Pan et al., 2005), reinforcement learning research has been the focal point of immense research progress (Mnih et al., 2015; Hasselt et al., 2016). Deep reinforcement learning has become an emerging field in the past decade with the introduction of deep neural networks as function approximators leading to learning policies that can surpass human cognitive abilities in highly complicated tasks by solely interacting with a given environment through trial and error without any supervision, consequently resulted in building AI agents that can reason and strategize (Mnih et al., 2015; Kapturowski et al., 2023; Krishnamurthy et al., 2024). In parallel, advances in neuroscience revealed the precise structures and neural circuitry dedicated to the computation of counterfactual state-action values in the human brain, and how these values are later compared to make decisions. In the specialized neural circuitry that underpins decision-making a compelling functional divide has been identified: while the prefrontal cortex encodes the expected values of executed actions, the dorsomedial frontal cortex plays a critical role in the analysis of counterfactual decisions providing the mechanisms for learning that can reason and generalize (Wunderlich et al., 2009; Lau & Glimcher, 2007; Klein-Flügge et al., 2016).

Beyond the initial inspiration by neuroscience, reinforcement learning further offers strong, mathematically provable, asymptotic guarantees on its ability to learn policies for solving complex problems via trial and error (Sutton, 1984; Watkins & Dayan, 1992). Nonetheless, a recent body of work exposed critical safety concerns of reinforcement learning, and consequently, a new class of algorithms has emerged that modify standard reinforcement learning algorithms to ensure reliability and safety in deep reinforcement learning (Madry et al., 2018; Korkmaz, 2024).

In this paper, we investigate the core intuition of reinforcement learning, and we analyze the theoretical underpinnings of counterfactuality and alignment in connection to the neuroscientific analysis of natural intelligence, and the consequential effects of trying to ensure safety in reinforcement learning. Our analysis discovers that the line of research focused on safety fails to deliver the guarantees implied by *certified safety and robustness*, and further risks potentially significant changes to the behavior and semantics of the trained policies that disrupts the foundations of reinforcement learning and its inherent capabilities. Essentially in this paper we aim to seek answers for the following questions: *(i) What are the consequences of current efforts to explicitly impose safety on reinforcement learning? (ii) What are the underlying reasons for preserving the core intuitive principles, neuroscientific foundations, and inherent capabilities of reinforcement learning?* To be able to answer these questions we focus on the foundations of reinforcement learning and its alignment with natural intelligence, and make the following contributions:

Contributions. We first provide a theoretically well-founded rigorous analysis of the state-action value function learnt by methods explicitly enforcing ϵ-local invariance and standard reinforcement learning in Section 3. Our analysis uncovers fundamental insights into how ϵ-local invariance imposition alters the very fabric of an agent's learned value judgments. Our paper is the first one that demonstrates, both theoretically and empirically, that methods explicitly enforcing robustness in fact fundamentally disrupt the inherent learning processes of standard reinforcement learning, consequentially leading to the subversion and loss of essential skills. Our analysis reveals that reinforcement learning possesses an inherent ability for counterfactual reasoning and is naturally aligned with human decision-making processes, while a recent line of work focusing on enforcing standard reinforcement learning to be explicitly robust causes standard RL policies to lose the inherent counterfactual ability and results in learning policies that are inaccurate, inconsistent and misaligned. We then conduct experiments in MDPs with high-dimensional state spaces from the Arcade Learning Environment (ALE) in Section 4, and our comprehensive study verifies the theoretical analysis and demonstrates a critical trade-off. Our findings reveal that standard deep neural policies naturally retain core skills that align with the value assignment of natural intelligence which allows them to reason and generalize. However, subjecting them to explicit ϵ-local invariance training shatters this elegant relationship and eradicates this intrinsic counterfactual ability and alignment. Our paper establishes the foundational principle of an intrinsic trade-off between counterfactuality and robustness and further uncovers the core mechanisms driving this fundamental trade-off as a direct result of certified training.

2 Background and Preliminaries

Markov Decision Process. An MDP is represented by a tuple $\mathcal{M} = (S, \mathcal{A}, P, r, \rho_0, \gamma)$ where S is a set of continuous states, $\mathcal{A}$ is a discrete set of actions, P is a transition probability distribution on $S \times \mathcal{A} \times S$, $r : S \times \mathcal{A} \to \mathbb{R}$ is a reward function, ρ_0 is the initial state distribution, and γ is the discount factor. The objective in reinforcement learning is to learn a policy $\pi : S \to P(\mathcal{A})$ which maps states to probability distributions on actions in order to maximize the expected cumulative reward $R = \mathbb{E} \sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)$ where $a_t \sim \pi(s_t)$. In $\mathcal{Q}$-learning (Watkins, 1989) the goal is to learn the optimal state-action value function $\mathcal{Q}^*(s, a) = R(s, a) + \sum_{s' \in S} P(s'|s, a) \max_{\hat{a} \in \mathcal{A}} \mathcal{Q}^*(s', \hat{a})$. Thus, the optimal policy is determined by choosing the action $a^*(s) = \arg\max_a \mathcal{Q}(s, a)$ in state s.

Adversarial Crafting and Training. Concerns regarding ϵ-invariance start with the work of Goodfellow et al. (2015), who observed that perturbations that are imperceptible to natural intelligence can in fact change the decision of a deep neural network and further suggested a fast method to produce such perturbations based on the linearization of the cost function used in training the network. Kurakin et al. (2016) proposed the iterative version of the fast gradient sign method proposed by Goodfellow et al. (2015) inside an ϵ-ball

$$x_{\text{adv}}^{N+1} = \text{clip}_\epsilon(x_{\text{adv}}^N + \alpha \text{sign}(\nabla_x J(x_{\text{adv}}^N, y))) \tag{1}$$

in which $J(x, y)$ represents the cost function used to train the deep neural network, x represents the input, and y represents the output labels. While several other methods have been proposed Korkmaz (2024) using a momentum-based extension of the iterative fast gradient sign method,

$$v_{t+1} = \mu \cdot v_t + \frac{\nabla_{s_{\text{adv}}} J(s_{\text{adv}}^t + \mu \cdot v_t, a)}{\|\nabla_{s_{\text{adv}}} J(s_{\text{adv}}^t + \mu \cdot v_t, a)\|_1} \quad , \quad s_{\text{adv}}^{t+1} = s_{\text{adv}}^t + \alpha \cdot \frac{v_{t+1}}{\|v_{t+1}\|_2}$$

robust training, i.e. ϵ-invariance training, has mostly been conducted with perturbations computed by projected gradient descent, i.e. PGD, proposed by Madry et al. (2018) (i.e. Equation 1).

Neuroscientific Results and Alignment with Natural Intelligence Decisions Making. A notable aspect of human cognitive decision-making is the assignment of counterfactual values to alternative, unchosen decisions (Wunderlich et al., 2009; Lee et al., 2012; Phillips et al., 2019). This mechanism serves to inform future decision-making by preserving a clear ordering of both factual and counterfactual outcomes and is a key attribute of the decision-making process that enables generalization and reasoning (Hoeck et al., 2015; Phillips et al., 2019; Grabenhorst & Rolls, 2011). Notably, the results in Figure 1 report analysis of fMRI scans of human brains during a decision-making task to identify a neural structure that compares the values of chosen and unchosen options for a particular decision. The results demonstrate that the value of each option was encoded in this structure, and that the actual decisions made were accurately predicted by these values (Klein-Flügge et al., 2016).

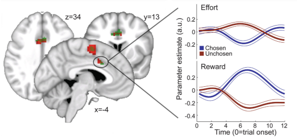

Figure 1: Human decision making and value assignment for options (Klein-Flügge et al., 2016).

Concerns on Reliability of Deep Neural Policies. The initial investigation on volatilities of deep neural policies was conducted based on testing ϵ-invariance via the utilization of the fast gradient sign method proposed by Goodfellow et al. (2015). Reliability of reinforcement learning policies has been further analyzed, and some studies argued the existence of shared ϵ-variant directions across MDPs and algorithms can be due to an underlying linear structure learnt by policies (Korkmaz, 2022; 2024). While several studies focused on improving optimization techniques for computing optimal perturbations, a line of research focused on making deep neural policies resilient to these perturbations. In particular, Pinto et al. (2017) proposed to model the dynamics between the adversary and the deep neural policy as a zero-sum game where the goal of the adversary is to minimize expected cumulative rewards of the deep neural policy. Gleave et al. (2020) approached this problem with an adversary model which is restricted to take natural actions in the MDP instead of modifying the observations with ℓ_p-norm bounded perturbations. The authors model this dynamic as a zero-sum Markov game and solve it via self-play. Recently, Huan et al. (2020) proposed to model this interaction as a state-adversarial MDP, and further claimed that their proposed algorithm SA-Double Deep Q-Network (SA-DDQN) learns theoretically certified ϵ-invariant policies against both natural noise and perturbations. Recent studies surprisingly revealed that certified training exhibits generalization issues with unpredictable behaviour and larger oscillations compared to standard reinforcement learning (Korkmaz, 2024). Despite these observations on generalization, currently still a large body of work is produced on explicitly optimizing variants of ϵ-invariance training, without any foundational analysis and explanation on how precisely this class of algorithms affect the inherent abilities of standard reinforcement learning and why this line of approach might not be the way of achieving true reliability.

3 The Core Intuition of Reinforcement Learning: The Inherent Counterfactuality

Our extensive analysis and results discover that ϵ-invariance training methods break the core intuitive principles of reinforcement learning and erode the inherent skills of RL policies up to the level of learning random values for counterfactual actions. Our analysis reveals that a key and strong attribute of reinforcement learning that allows generalization and reasoning is lost when subjected to invariance training. Our study provides evidence that the application of ϵ-invariance training fails to fully address the critical issues of robustness and safety in modern AI. The observed misalignment between this training methodology and desirable system behavior reveals a key dichotomy: while certified training aims for provable guarantees, it appears to diverge from the inherent principles that underpin the core attributes of natural intelligence with regard to reasoning and generalization.

The theoretically motivated ϵ-locally invariant (ϵ-LI) training techniques achieve certified defense against perturbations inside the ϵ-ball, $\mathcal{D}_\epsilon(s) = \{\bar{s} : \|s - \bar{s}\|_\infty \leq \epsilon\}$, forming the current foundations of the robust reinforcement learning. However, we provide foundational evidence that this approach induces significant changes in the $\mathcal{Q}$-function where the state-action value function no longer accurately represents the MDP. In particular, ϵ-locally invariant (ϵ-LI) training causes deep neural policies to learn misaligned, inaccurate, overestimated state-action value functions while causing standard reinforcement learning to lose it's inherent counterfactuality. Furthermore, we connect and highlight the neural processing of decision making of natural intelligence, core intuition of reinforcement learning and certified training (Wunderlich et al., 2009; Lau & Glimcher, 2007; Grabenhorst & Rolls, 2011). Our results reveal that certified training disrupts the core intuition of reinforcement learning and leads to learning policies that are disjoint and orthogonal to natural intelligence decision making, i.e. the true biological inspiration of reinforcement learning (Romo & Schultz, 1990; Montague et al., 1996). In the remainder of this section we will provide the theoretical foundations on: *i. Why we need to preserve the core intuition of reinforcement learning*, and *ii. What factors precisely disrupts the essential core skills learnt by reinforcement learning*. The theoretical underpinning of ϵ-invariance training methods is derived from Danskin's theorem.

Theorem 3.1 (Danskin (1967))**.** *Let $\mathcal{X}$ be a compact topological space $f : \mathbb{R}^n \times X \to \mathbb{R}$, $f(\cdot, x)$ is differentiable for every $x \in \mathcal{X}$, $x^*(\theta) = \{x \in \arg\max_{x\in\mathcal{X}} f(\theta, x)\}$ and $\nabla_\theta f(\theta, x)$ is continuous on $\mathbb{R}^n \times \mathcal{X}$. Then the max function $\kappa(\theta) = \max_{x\in\mathcal{X}} f(\theta, x)$ is locally Lipschitz continuous, directionally differentiable, and its directional derivatives satisfy $\kappa'(\theta, h) = \sup_{x\in x^*(\theta)} h^\top \nabla_\theta f(x, \theta)$. Furthermore, if the set $x^*(\theta)$ has size one i.e. there is a unique maximizer x^*_θ then $\nabla_\theta \kappa(\theta) = \nabla_\theta f(\theta, x^*_\theta)$.*

Danskin's theorem provides a method to compute the gradient of a function that is defined in terms of a maximization over a set. With this theoretically well-motivated start, a line of algorithms have been proposed to make models reliable. The approach of ϵ-invariance training techniques is based on editing the standard $\mathcal{Q}$-learning update. This change made to the update is designed to penalize $\mathcal{Q}$-functions for which a perturbed state $\bar{s} \in \mathcal{D}_\epsilon(s)$ can change the identity of the highest $\mathcal{Q}$-value action. Formally, the canonical definition used in the literature is

Definition 3.2 (*Robust reinforcement learning*)**.** Within an ϵ-neighbourhood the reinforcement learning policy should be invariant to $\mathcal{D}_\epsilon(s) := \{\hat{s} \in S \mid \hat{s} \in \mathcal{D}_\epsilon(s), \text{argmax}_a \mathcal{Q}(s, a) = \text{argmax}_a \mathcal{Q}(\hat{s}, a)\}$ where $\mathcal{D}_\epsilon(s) = \{\hat{s} : \|s - \hat{s}\|_p \leq \epsilon\}$. Then the policy is ϵ-invariant (robust).

Now we will prove that there is a fundamental trade-off between accurate estimation of $\mathcal{Q}$-values and robustness. In particular, the optimal state-action value function $\mathcal{Q}^*$ is not ϵ-invariant, but there is a ϵ-invariant state-action value function $\mathcal{Q}_\theta$ that overestimates the optimal state-action values.

Theorem 3.3 (*Inherent trade-off between estimation and robustness*)**.** *Let $\epsilon > 0$. In the linear function approximation setting, there is an MDP such that all linear-state action value functions matching the optimal state-action values $\mathcal{Q}^*$ are not ϵ-invariant. Furthermore, there is a linear state-action value function $\mathcal{Q}_\theta$ that is ϵ-invariant, but overestimates the optimal state-action values while maintaining the correct optimal action.*

Proof. Let there be two states s_1 and s_2 such that $\|s_1 - s_2\|_2 = 1$. Further suppose that the optimal state-action values satisfy $\mathcal{Q}^*(s_1, a_1) = \epsilon/10$, $\mathcal{Q}^*(s_1, a_2) = 0$, $\mathcal{Q}^*(s_2, a_1) = 0.8$, and $\mathcal{Q}^*(s_2, a_2) = 1.0$. Next let $\mathcal{Q}_\theta(s, a)$ be any linearly parameterized state-action value function that agrees with $\mathcal{Q}^*(s, a)$ on the states s_1 and s_2. Consider the one-dimensional functions $\Psi_1(\xi) = \mathcal{Q}_\theta((1 - \xi) \cdot s_1 + \xi \cdot s_2, a_1)$ and $\Psi_2(\xi) = \mathcal{Q}_\theta((1 - \xi) \cdot s_1 + \xi \cdot s_2, a_2)$ which are the restriction of $\mathcal{Q}_\theta(s, a)$ to the line segment from s_1 to s_2. By linearity of $\mathcal{Q}_\theta$ we also have that both Ψ_1 and Ψ_2 are linear. Furthermore, since $\mathcal{Q}_\theta$ agrees with $\mathcal{Q}^*$ at s_1 and s_2, we know the values of both functions at two points i.e. $\Psi_1(0) = \mathcal{Q}^*(s_1, a_1)$, $\Psi_1(1) = \mathcal{Q}^*(s_2, a_1)$, $\Psi_2(0) = \mathcal{Q}^*(s_1, a_2)$, and $\Psi_2(1) = \mathcal{Q}^*(s_2, a_2)$. As Ψ_1 and Ψ_2 are linear functions on $\mathbb{R}$, the values at two points are sufficient to uniquely determine the functions. In particular we have

$$\Psi_1(\xi) = (0.8 - \epsilon/10)\xi + \epsilon/10 \quad \text{and} \quad \Psi_2(\xi) = \xi$$

Note that these two lines intersect at the point $\hat{\xi} = \frac{\epsilon}{2+\epsilon}$. Let $\hat{s} = (1-\hat{\xi}) \cdot s_1 + \hat{\xi} \cdot s_2$. Since the lines of Ψ_1 and Ψ_2 intersect at $\hat{\xi}$, we conclude that $\mathcal{Q}_\theta(\hat{s}, a_2) \geq \mathcal{Q}_\theta(\hat{s}, a_1)$. However, $\mathcal{Q}_\theta(s_1, a_1) > \mathcal{Q}_\theta(s_1, a_2)$. Furthermore, $\|s_1 - \hat{s}\| = \frac{\epsilon}{2+\epsilon} < \epsilon$. Thus, $\mathcal{Q}_\theta$ is not ϵ-invariant. However, if we instead choose new parameters θ' for the state-action value function so that $\mathcal{Q}_{\theta'}(s_1, a_1) = 0.8$ and $\mathcal{Q}_{\theta'}(s_1, a_2) = 0.7$ one can easily check that $\mathcal{Q}_{\theta'}$ is ϵ-invariant for all $\epsilon < 0.1$. Furthermore, observe that $\mathcal{Q}_{\theta'}$ gives the correct ranking of actions in state s_1, but overestimates the optimal state-action value by a factor of $8/\epsilon$. □

The results reported in Section 4 verify the theoretical analysis on the fundamental trade-off between estimation and ϵ-invariance in neural-network approximation of the $\mathcal{Q}$-function. Now we will further theoretically analyze the effects of canonical ϵ-invariance training techniques (Huan et al., 2020).

Definition 3.4 (*Baseline Certified ϵ-invariance Training*)**.** The regularizer that achieves certified ϵ-invariance inside the ϵ-ball $\mathcal{D}_\epsilon(s) = \{\bar{s} : \|s - \bar{s}\|_\infty \leq \epsilon\}$ for $\mathcal{Q}_\theta(s,a)$ $\forall \bar{s} \in \mathcal{D}_\epsilon(s)$ is

$$\mathcal{R}(\theta) = \sum_s \Big(\max_{\bar{s}\in\mathcal{D}_\epsilon(s)} \max_{a \neq \arg\max_a \mathcal{Q}_\theta(s,a)} \mathcal{Q}_\theta(\bar{s},a) - \mathcal{Q}_\theta(\bar{s}, \arg\max_a \mathcal{Q}_\theta(s,a))).$$

The certified training algorithm proceeds by adding $\mathcal{R}(\theta)$ to the standard temporal difference loss $\mathcal{L}_\mathcal{H}\left(r(s,a) + \gamma \max_{a'} \mathcal{Q}^{\text{target}}(s',a') - \mathcal{Q}_\theta(s,a)\right) + \mathcal{R}(\theta)$.

Now we will show that changing the standard $\mathcal{Q}$-update will cause losing counterfactuality $\forall a \in \mathcal{A}_s^\perp$ where $\mathcal{A}_s^\perp := \{a | a \neq \arg\max_{\hat{a}} \mathcal{Q}(s,\hat{a})\}$, and overestimation of the state-action values $\forall a \in \mathcal{A}$. For this now let us look at the MDP $\mathcal{M}$ where two states parametrized by feature vectors $s_1, s_2 \in \mathbb{R}^n$, with three possible actions $\{a_i\}_{i=1}^3$ in each state where taking any of the actions in state s_1 leading to a transition to state s_2 and vice versa. Let $1 > \gamma > 0$ be the discount factor, and let $\delta > \eta > 0$ be small constants with $\gamma > \delta$. The rewards for each action are as follows: $r(s_1,a_1) = 1-\gamma$, $r(s_1,a_2) = \eta-\gamma$, $r(s_1,a_3) = \delta - \gamma$, $r(s_2,a_1) = \eta - \gamma$, $r(s_2,a_2) = 1 - \gamma$, and $r(s_2,a_3) = \delta - \gamma$. Clearly, the optimal policy is to always take action a_1 in state s_1, and action a_2 in state s_2 as these are the only actions giving positive reward. Thus the optimal state-action values are given by: $\mathcal{Q}^*(s_1,a_1) = \mathcal{Q}^*(s_2,a_2) = \sum_{t=0}^\infty (1-\gamma)\gamma^t = 1$, $\mathcal{Q}^*(s_1,a_2) = \mathcal{Q}^*(s_2,a_1) = \eta - \gamma + \gamma \sum_{t=0}^\infty (1-\gamma)\gamma^t = \eta$, and $\mathcal{Q}^*(s_1,a_3) = \mathcal{Q}^*(s_2,a_3) = \delta - \gamma + \gamma\sum_{t=0}^\infty (1-\gamma)\gamma^t = \delta$. Let the $\mathcal{Q}$-function be linearly parametrized by $\theta = (\theta_1, \theta_2, \theta_3)$ so that $\mathcal{Q}_\theta(s,a_i) = \langle \theta_i, s\rangle$. Finally, let Φ_i for $i \in \{1,2,3\}$ be three orthonormal vectors, and let the state feature vectors satisfy:

$$1.\ s_1 = \Phi_1 + \delta\Phi_3 + \eta\Phi_2 \quad \text{and} \quad 2.\ s_2 = \Phi_2 + \delta\Phi_3 + \eta\Phi_1$$

Then it follows that the optimal $\mathcal{Q}$-function is parametrized by $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*)$ where $\theta_i^* = \Phi_i$ i.e. $\mathcal{Q}_{\theta^*}(s,a) = \mathcal{Q}^*(s,a)$ for all s and a. Thus, according to the function $\mathcal{Q}_{\theta^*}(s,a)$, for s_1 the best action is a_1, for s_2 the best action is a_2, and in all states the second-best action is a_3. Next we identify the optimal perturbations used in the computation of the regularizer $\mathcal{R}(\theta^*)$ for this setting.

Proposition 3.5. *In the MDP $\mathcal{M}$ for any $\epsilon > 0$.*

1. *For $s = s_1$:* $s + \frac{\epsilon}{\sqrt{2}}(\theta_3^* - \theta_1^*) = \arg\max_{\bar{s}\in\mathcal{D}_\epsilon(s)} \max_{a\neq a^*(s)} \mathcal{Q}_{\theta^*}(\bar{s},a) - \mathcal{Q}_{\theta^*}(\bar{s},a^*(s))$
2. *For $s = s_2$:* $s + \frac{\epsilon}{\sqrt{2}}(\theta_3^* - \theta_2^*) = \arg\max_{\bar{s}\in\mathcal{D}_\epsilon(s)} \max_{a\neq a^*(s)} \mathcal{Q}_{\theta^*}(\bar{s},a) - \mathcal{Q}_{\theta^*}(\bar{s},a^*(s))$

Proof. We will prove item 1, and item 2 will follow from an identical argument with roles of θ_1^* and θ_2^* swapped. Let $s = s_1$. Since $a^*(s) = 1$, there are two case to consider for the maximum over $a \neq a^*(s)$, either $a = 2$ or $a = 3$. In the case that $a = 2$ we have

$$\max_{\bar{s}\in\mathcal{D}_\epsilon(s)} \mathcal{Q}_{\theta^*}(\bar{s},a) - \mathcal{Q}_{\theta^*}(\bar{s},a^*(s)) = \max_{\bar{s}\in\mathcal{D}_\epsilon(s)} \langle\theta_2^*, \bar{s}\rangle - \langle\theta_1^*, \bar{s}\rangle. \tag{2}$$

This is the maximum in a ball of radius ϵ around s of the linear function $\langle\theta_2^* - \theta_1^*, \bar{s}\rangle$. Therefore the maximum is achieved by $\bar{s} = s + \frac{\epsilon}{\sqrt{2}}(\theta_2^* - \theta_1^*)$. The corresponding maximum value is

$$\max_{\bar{s}\in\mathcal{D}_\epsilon(s)} \langle\theta_2^*, \bar{s}\rangle - \langle\theta_1^*, \bar{s}\rangle = \langle\theta_2^* - \theta_1^*, s\rangle + \epsilon\|\theta_2^* - \theta_1^*\|_2 = \eta - 1 + \epsilon\sqrt{2}. \tag{3}$$

In the case that $a = 3$ an identical argument implies that the maximum is achieved by $\bar{s} = s + \frac{\epsilon}{\sqrt{2}}(\theta_3^* - \theta_1^*)$, with corresponding maximum value

$$\max_{\bar{s}\in\mathcal{D}_\epsilon(s)} \langle\theta_3^*, \bar{s}\rangle - \langle\theta_1^*, \bar{s}\rangle = \langle\theta_3^* - \theta_1^*, s\rangle + \epsilon\|\theta_3^* - \theta_1^*\|_2 = \delta - 1 + \epsilon\sqrt{2}. \tag{4}$$

Because $\delta > \eta$ we conclude that the value achieved in 4 is larger than that in 3. Thus the maximizer is $\bar{s} = s + \frac{\epsilon}{\sqrt{2}}(\theta_3^* - \theta_1^*)$ as desired. □

In words, the optimal direction to perturb the state s_1 in order to have $a^*(s) \neq a^*(\bar{s})$ is toward $\theta_3^* - \theta_1^*$. Similarly for the state s_2, the optimal perturbation is toward $\theta_3^* - \theta_2^*$. Next we use this fact to show that in order to decrease the regularizer it is sufficient to simply increase the magnitude of θ_1 and θ_2, and decrease the magnitude of θ_3.

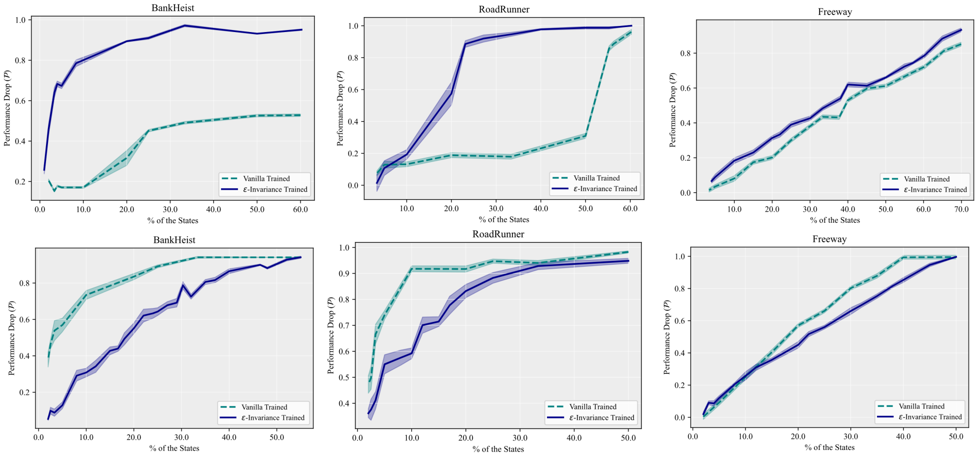

Figure 2: Up: Performance drop $\mathcal{P}_2(\Omega)$ with respect to action modification a_2 for the state-of-the-art ϵ-invariance and vanilla trained deep neural policies. Down: Performance drop $\mathcal{P}_w(\Omega)$ with respect to action modification a_w. Left: BankHeist. Center: RoadRunner. Right: Freeway.

Proposition 3.6. *In the MDP $\mathcal{M}$ let $\lambda > 0$ and suppose that $(1 - \lambda)\delta < (1 + \lambda)\eta < \delta$. Let $\theta = (\theta_1, \theta_2, \theta_3)$ be given by $\theta_1 = (1 + \lambda)\theta_1^*$, $\theta_2 = (1 + \lambda)\theta_2^*$ and $\theta_3 = (1 - \lambda)\theta_3^*$. Then $\mathcal{R}(\theta) < \mathcal{R}(\theta^*)$.*

The proof of Proposition 3.6 is provided in the supplementary material. Combining Proposition 3.6 and Proposition 3.5 we can prove the main result of this section on the effects ϵ-invariance training.

Theorem 3.7 (*Existence of Overestimation and Misalignment of Counterfactual Decisions*)**.** *There is an MDP with linearly parameterized state-action values, optimal state-action value parameters θ^*, and a parameter vector θ such that: $\mathcal{L}(\theta) < \mathcal{L}(\theta^*)$, and the parameter vector θ overestimates the optimal state-action value and re-orders the sub-optimal ones.*

The proof of Theorem 3.7 is provided in the supplementary material. The results reported in Section 4 verify the fundamental trade-off and the theoretical predictions of Section 3. In particular, across a diverse portfolio of state-of-the-art ϵ-invariance training techniques that aim to obtain safe and reliable reinforcement learning, our results demonstrate that certified ϵ-invariance trained policies learn misaligned, inacccurate and inconsistent values while further losing counterfactuality compared to standard reinforcement learning.

4 Emprical Analysis in High-Dimensional MDPs

The empirical analysis is conducted in high dimensional state representation MDPs of the Arcade Learning Environment (ALE). The standard reinforcement learning policy is trained via DDQN (Wang et al., 2016) initially proposed in (van Hasselt, 2010) with prioritized experience replay proposed by (Schaul et al., 2016), and the ϵ-invariance reinforcement learning policies are trained via SA-MDP RL (State Adversarial MDP, see Section 2), RADIAL (Robust Adversarial Loss-RL), and Optimal Robust Policy (ORP) (Li et al., 2024) where all of these influential studies were attributed as oral and spotlight presentations at NeurIPS and ICML respectively. The standard error of the mean is included for all of the figures and tables. See supplementary material for the hyperparameters and the implementation details. Performance drop $\mathcal{P}$ is given by $\mathcal{P} = (\text{Score}_{\text{base}} - \text{Score}_{\text{actmod}})/(\text{Score}_{\text{base}} - \text{Score}_{\text{min}})$, where $\text{Score}_{\text{base}}$ represent the baseline run of the game without modification, $\text{Score}_{\text{min}}$ represents the minimum score available for a given game, and $\text{Score}_{\text{actmod}}$ represents the run of the game where the actions of the agent are modified for a fraction of the state observations. To measure the accuracy for the state-action value estimates formally, let a_i be the i^{th} best action decided by the deep neural policy in a given state s (i.e. $\mathcal{Q}(s, a)$ is sorted in decreasing order, and a_i is the action corresponding to i^{th} largest $\mathcal{Q}$-value). For a trained agent, the value of $\mathcal{Q}(s, a_i)$ should represent the expected cumulative rewards obtained by taking action a_i in state s, and then taking the highest $\mathcal{Q}$-value action (i.e. a_1) in every subsequent state. Thus, a natural test to perform would be: for a random state s the policy should take action a_i in state s, and the highest $\mathcal{Q}$-value action for the rest of the states. By comparing the relative performance drop $\mathcal{P}$ in this test to a clean run where the agent always takes the highest $\mathcal{Q}$-value action, one can measure the decline in rewards caused by taking action a_i. Further, we can provide a measure of accuracy for the state-action value function by comparing the results of the test

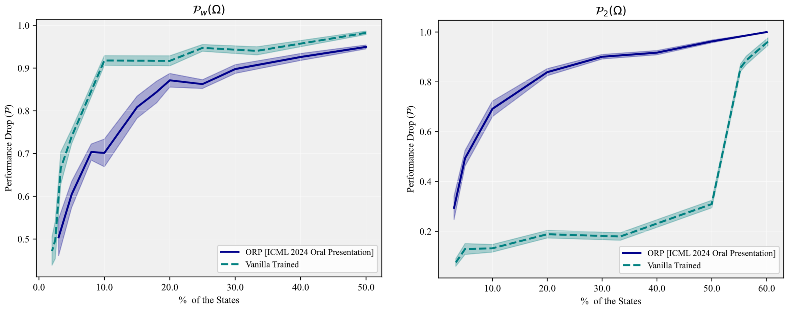

Figure 3: Loss of counterfactuality: $\mathcal{P}_2(\Omega)$ and $\mathcal{P}_w(\Omega)$ results with respect to a_2 and a_w for ORP (Optimal Robust Policy) ϵ-invariance reinforcement learning and vanilla reinforcement learning.

for each $i \in \{1, 2 \dots |A|\}$, and checking that the relative performance drops $\mathcal{P}_i$ are in the correct order i.e. $0 = \mathcal{P}_1 \leq \mathcal{P}_2 \dots \leq \mathcal{P}_{|A|}$. We take this one step further and analyze the performance drop with Ω-fraction of the states in the episode uniformly at random, and making the policy execute action a_i in each of the sampled states. We then record the relative performance drop as a function of Ω, yielding a performance drop curve $\mathcal{P}_i(\Omega)$. More formally, the performance curve is

Definition 4.1 (*Performance Drop Curve*). Let $\mathcal{M}$ be an MDP and $\mathcal{Q}(s, a)$ be a state-action value function for $\mathcal{M}$. In each state label the actions $a_1, \dots a_{|A|}$ in order so that $\mathcal{Q}(s, a_1) \geq \mathcal{Q}(s, a_2) \dots \geq \mathcal{Q}(s, a_{|A|})$. The *performance drop curve* $\mathcal{P}_i(\Omega)$ is the expected performance drop of an agent in $\mathcal{M}$ which takes action a_i in a randomly sampled Ω-fraction of states, and executes a_1 in all other states.

Using these performance drop curves one can confirm whether $\mathcal{P}_i(\Omega)$ lies above $\mathcal{P}_j(\Omega)$ whenever $i > j$. Yet to be precise we will quantify the relative ordering of the performance drop curves.

Definition 4.2 (*τ-domination*). Let $\mathcal{F} : [0, 1] \to [0, 1]$ and $\mathcal{G} : [0, 1] \to [0, 1]$. For any $\tau > 0$, we say that the $\mathcal{F}$ τ-dominates $\mathcal{G}$ if $\int_0^1 (\mathcal{F}(\Omega) - \mathcal{G}(\Omega)) \, d\Omega > \tau$.

To compare the accuracy of state-action values for vanilla versus ϵ-invariance trained agents, we can thus perform the above test, and check the relative ordering of the curves $\mathcal{P}_i(\Omega)$ using Definition 4.2 for each agent type. In addition, we can also directly compare for each i the curve $\mathcal{P}_i^{\text{adv}}(\Omega)$ for the ϵ-invariance trained agent with the curve $\mathcal{P}_i^{\text{vanilla}}(\Omega)$ of the vanilla trained agent. This is possible because $\mathcal{P}_i(\Omega)$ measures the performance drop of the agent relative to a clean run, and thus always takes values on a normalized scale from 0 to 1. Hence, an observation of $\mathcal{P}_2^{\text{adv}}(\Omega)$ τ-dominating $\mathcal{P}_2^{\text{vanilla}}(\Omega)$ for some $\tau > 0$, this would conclude that the state-action value function of the vanilla trained agent can accurately represent the counterfactual actions than the ϵ-invariance trained agents.

4.1 Losing Inherent Counterfactuality

In Section 3 we provided theoretical analysis on how ϵ-invariance training effects the core principles of reinforcement learning. In this section, we demonstrate that standard reinforcement learning is inherently counterfactual and certified ϵ-invariance training causes the policy to lose counterfactuality. Figure 2 and Figure 3 report the performance drop $\mathcal{P}_2(\Omega)$ and $\mathcal{P}_w(\Omega)$ as a function of the fraction of states Ω in which the action modification is applied for ϵ-invariance and vanilla trained deep neural policies. In particular, the action modification is set for the second best action a_2 decided by the state-action value function $\mathcal{Q}(s, a)$. As the fraction of states for $\mathcal{P}_2(\Omega)$ increases, vanilla trained deep neural policies experience lower performance drops compared to ϵ-invariance. Especially in BankHeist we observe that the performance drop does not exceed 0.55 even when the action modification is applied for a large fraction of the visited states for the standard reinforcement learning policies. This gap in the performance drop between the ϵ-invariance and vanilla trained deep neural policies indicates that the state-action value function learnt by standard reinforcement learning has a better estimate for the state-action values. We further investigate the effects of $a_w = \arg\min_a \mathcal{Q}(s, a)$, i.e. worst possible action in a given state, modification on the deep neural policy. Intriguingly, Figure 2 and 3 report that the performance drop $\mathcal{P}_w(\Omega)$ is higher in the vanilla trained deep neural policies compared to ϵ-invariance trained ones when the action modification is set to a_w. This again further verifies the theoretical predictions in Section 3 and demonstrates that standard reinforcement learning learns a counterfactual and accurate state-action value function while ϵ-invariance policies lose core inherent skills of reinforcement learning.

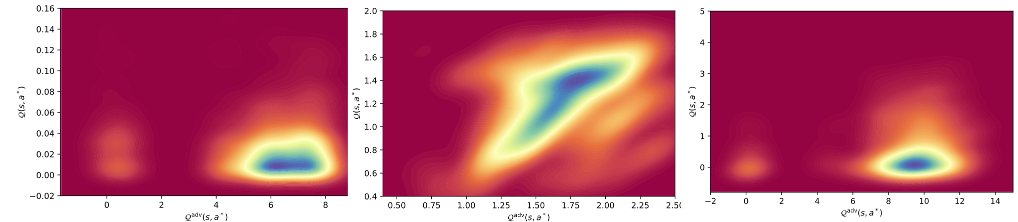

Figure 4: $\mathcal{Q}$ values of $\arg\max_{a\in\mathcal{A}} \mathcal{Q}(s,a)$ for ϵ-invariance and vanilla trained deep neural policies.

MDPs	BankHeist		RoadRunner		Freeway	
Method	ϵ-Invariance	Vanilla	ϵ-Invariance	Vanilla	ϵ-Invariance	Vanilla
AM a_2	0.449±0.007	**0.191±0.04**	0.414±0.015	**0.247±0.009**	0.351±0.009	**0.302±0.007**
AM a_w	0.311± 0.011	**0.398±0.011**	0.345±0.011	**0.393±0.002**	0.241±0.007	**0.311±0.010**

Table 1: Area under the curve of performance drop under action modification (AM) a_2 and a_w for the state-of-the-art ϵ-invariance trained deep neural policies and vanilla trained deep neural policies.

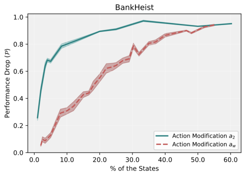

Figure 5: $\mathcal{P}_2$ and $\mathcal{P}_w$ of certified ϵ-invariance training.

Reinforcement learning has inherent counterfactual ability and intrinsically learns aligned values.

The progression of AI, from foundational work in perception to advanced decision-making, has been marked by key milestones driven by concepts drawn from biological inspiration (Treisman & Gelade, 1980; Snowden et al., 1991; Rao & Ballard, 1999; Parthasarathy et al., 2024; Newell, 1992; Imaizumi et al., 2022; Hassabis et al., 2017) Reinforcement learning is founded on the inspiration drawn from natural intelligence (ichi Amari & Arbib, 1982; Kehoe et al., 1987; Romo & Schultz, 1990; Montague et al., 1996) providing further theoretical guarantees on its limitations and capabilities (Watkins & Dayan, 1992; Barto et al., 1995). Our results show that ϵ-invariance training compromises the foundational intuition of reinforcement learning, leading to a loss of the inherent counterfactuality and creating significant value misalignment. Our analysis and results demonstrate that an extensive recent line of work myopically focusing on safety in fact diverts the main contributions and the tight core connection of reinforcement learning with neuroscience while producing policies that are both in fact unsafe and misaligned. In particular, Figure 5 demonstrates that choosing the worst action leads to a smaller performance drop than choosing the second best action i.e. $\mathcal{P}_w(\Omega) < \mathcal{P}_2(\Omega)$ for all Ω. Notably, these results reveal that ϵ-invariance training methods assign random values to the counterfactual actions which is a direct misalignment with natural intelligence decision making. The results reported in Figure 2 demonstrate the clear juxtaposition between standard reinforcement learning and reliability-concerned reinforcement learning, i.e. ϵ-invariance. Intriguingly, these findings reveal that standard reinforcement learning successfully learns aligned values and possesses an inherent capacity for counterfactual reasoning. Imposing reinforcement learning to be ϵ-invariant strips out these intrinsic skills. While learning inconsistent and misaligned values can cause vulnerability problems from a security point of view, our analysis further highlights the foundational loss of information in the state-action value function as a novel fundamental trade-off intrinsic to ϵ-invariance training.

Imposing ϵ-invariance causes misalignment and the loss of the inherent counterfactuality of RL.

Biased $\mathcal{Q}$-values in ϵ-invariance Trained Deep Neural Policies. In this section we investigate state-action value estimates of ϵ-invariance trained and vanilla trained deep neural policy. The results demonstrate that ϵ-invariance training leads to overestimation in $\mathcal{Q}$-values which verifies the theoretical analysis provided in Section 3. In particular, Figure 4 reports the overestimation bias on the state-action values learned by the ϵ-invariance trained deep neural policies. Note that the fact that ϵ-invariance trained policies assign higher state-action values than the vanilla trained deep reinforcement learning policies while performing similarly, i.e. obtaining similar expected cumulative rewards, reveals that the ϵ-invariance training techniques, on top of the misalignment, counterfactuality and the inaccuracy issues, learn explicitly biased state-action values.

Action Gap Phenomenon. The action gap is defined as the difference $\mathcal{Q}$-values, i.e. $\mathcal{G}(\mathcal{Q},s) = \max_{\hat{a}\in\mathcal{A}} \mathcal{Q}(s,\hat{a}) - \max_{a\in\mathcal{A}_s^\perp} \mathcal{Q}(s,a)$. A connection between the action gap and the approximation errors has been discussed in prior studies and it has been hypothesized that increasing the action gap

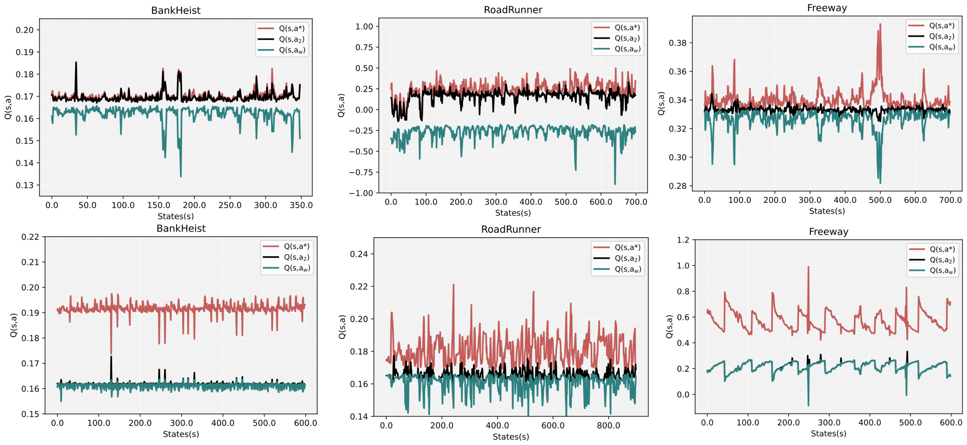

Figure 6: Normalized state-action values for the best action a^*, second best action a_2 and worst action a_w over states. Up: Vanilla trained. Down: State-of-the-art Lipshitz trained[2].

of the learned value function causes a decrease in overestimation of $\mathcal{Q}$-values. Following this study, several papers built on the hypothesis that increasing the action gap causes reduction in bias. However, our results reveal that targeting to increase the action gap must be upper-bounded by preserving the order of the counterfactual actions to obtain truly reliable and safe policies. Once this upperbound is passed the policy forms values that are misaligned and without the ability to think counterfactual. To preserve the core principles of reinforcement learning and its neuroscientific foundations that allow them to reason and generalize, we must preserve the approaches that targeted learning methods that align and match the foundational inspiration of reinforcement learning (Baird & Moore, 1993; Watkins & Dayan, 1992; Averbeck & Costa, 2017).

A Transparent Discussion and Call for Reconsideration. While these certified training algorithms have attracted a significant level of attention from the research community, including several spotlight and oral presentations at NeurIPS and ICML, to encourage more efforts on this line of research commiting to development of responsible policies, it is more significant than ever to discuss principled investigation of these approaches. If these issues are not openly and transparently discussed, it will harm the progress towards achieving true reliability and safety while influencing future research directions and significantly pivoting research efforts (Ren et al., 2024). Without the principled knowledge of the actual costs and drawbacks of these algorithms a significant level of research effort might be misdirected. The results reported in Section 4.1 and Figure 5, reveal concrete problems of the ϵ-invariance training techniques and how they erode reinforcement learning core skills including inherent counterfactuality and generalization. Our results call for an urgent reconsideration ϵ-invariance training of reinforcement learning and what constitutes true robustness.

5 Conclusion

In this paper, we focus on the core principles of reinforcement learning and how the inherent capabilities of RL policies are impacted by the efforts on explicit imposition of robustness. We provide an extensive theoretical analysis on the fundamental effects of ϵ-invariance training of reinforcement learning. Both our theoretical analysis and empirical analysis conducted in high-dimensional state representation MDPs reveal that standard reinforcement learning is inherently counterfactual and aligned with the human decision making process, while techniques focused on imposing ϵ-invariance erodes core skills of reinforcement learning. Moreover, our theoretical analysis reveals that there is a fundamental trade-off in ϵ-invariance training methods, and our empirical results demonstrate ϵ-invariance training breaks the core principles of reinforcement learning and causes policies to lose counterfactuality and learn misaligned and inaccurate state-action value functions. Our paper highlights transparent progress and calls for reconsideration of *robustness*, and our analysis is critical in understanding the true capabilities of standard reinforcement learning, and opens an avenue for more principled approach for designing algorithms to improve reliability.

[2] Figure 6 reports that ϵ-invariance training increases the action gap, yet still learns biased values. Due to the fact that the ϵ-invariance trained policy has biased $\mathcal{Q}$-values, the results are reported in the normalized form in order to compare the action gaps of ϵ-invariance and vanilla trained policies in the same graph.

REFERENCES

Bruno B Averbeck and Vincent D Costa. Motivational neural circuits underlying reinforcement learning. In *Nature Neuroscience*, 2017.

Leemon Baird and Andrew Moore. Reinforcement learning through gradient descent. In *Conference on Neural Information Processing Systems, NeurIPS*, 1993.

Andrew G. Barto, Steven J. Bradtke, and Satinder P. Singh. Learning to act using real-time dynamic programming. In *Artificial Intelligence*, 1995.

John M Danskin. A method for solving a convex programming problem with convergence rate $o(1/k^2)$. New York: Springer, 1967.

Adam Gleave, Michael Dennis, Cody Wild, Kant Neel, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. *International Conference on Learning Representations ICLR*, 2020.

Ian Goodfellow, Jonathan Shelens, and Christian Szegedy. Explaning and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

Fabian Grabenhorst and Edmund Rolls. Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences*, 2011.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. Neuron, 2017.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. *In Thirtieth AAAI conference on artificial intelligence*, 2016.

Nicole Van Hoeck, Patrick D. Watson, and Aron K. Barbey. Cognitive neuroscience of human counterfactual reasoning. Frontiers in Human Neuroscience 2015.

Zhang Huan, Chen Hongge, Xiao Chaowei, Bo Li, Mingyan Boning, Duane Liu, and ChoJui Hsiesh. Robust deep reinforcement learning against adversarial perturbations on state observatons. *NeurIPS Spotlight Presentation*, 2020.

Shun ichi Amari and Michael A. Arbib. Competition and cooperation in neural nets. Systems neuroscience, 1982.

Yuri Imaizumi, Agnieszka Tymula, Yasuhiro Tsubo, Masayuki Matsumoto, and Hiroshi Yamada. A neuronal prospect theory model in the brain reward circuitry. Nature, 2022.

Steven Kapturowski, Victor Campos, Ray Jiang, Nemanja Rakicevic, Hado van Hasselt, Charles Blundell, and Adrià Puigdomènech Badia. Human-level atari 200x faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023.

E. James Kehoe, Bernard G. Schreurs, and Peita Graham. Temporal primacy overrides prior training in serial compound conditioning of the rabbit's nictitating membrane response. Animal Learning and Behavior, 1987.

Miriam Klein-Flügge, Steven W. Kennerley, Karl Friston, and Sven Bestmann. Neural signatures of value comparison in human cingulate cortex during decisions requiring an effort-reward trade-off. In *The Journal of Neuroscience*, 2016.

Ezgi Korkmaz. Deep reinforcement learning policies learn shared adversarial features across mdps. *AAAI Conference on Artificial Intelligence*, 2022.

Ezgi Korkmaz. Understanding and Diagnosing Deep Reinforcement Learning Decision Making. In *International Conference on Machine Learning, ICML 2024*, 2024.

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Brian Lau and Paul W. Glimcher. Action and outcome encoding in the primate caudate nucleus. In *Journal of Neuroscience*, volume 27, pp. 14502–14514, 2007.

Daeyeol Lee, Hyojung Seo, and Min Whan Jung. Neural basis of reinforcement learning and decision making. The Annual Review of Neuroscience, 2012.

Haoran Li, Zicheng Zhang, Wang Luo, Congying Han, Yudong Hu, Tiande Guo, and Shichen Liao. Towards optimal adversarial robust q-learning with bellman infinity-error. In *Forty-first International Conference on Machine Learning, ICML 2024, [Oral Presentation]*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, arc G Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518: 529–533, 2015.

Read Montague, Peter Dayan, and Terrence Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. Journal of Neuroscience, 1996.

Allen Newell. Unified theories of cognition. Harvard University Press, 1992.

Wei-Xing Pan, Robert Schmidt, Jeffery R. Wickens, and Brian I. Hyland. Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. Journal of Neuroscience, 2005.

Nikhil Parthasarathy, Olivier J. Hénaff, and Eero P. Simoncelli. Layerwise complexity-matched learning yields an improved model of cortical area V2. *Transactions in Machine Learning Research*, 2024.

Jonathan Phillips, Adam Morris, and Fiery Cushman. How we know what not to think. Trends in Cognitive Sciences 2019.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. *International Conference on Learning Representations ICLR*, 2017.

Rajesh Rao and Dana Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 1999.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do AI safety benchmarks actually measure safety progress? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Ranulfo Romo and Wolfram Schultz. Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. Journal of Neurophysiology, 1990.

Tom Schaul, John Quan, Ioannis Antonogloua, and David Silver. Prioritized experience replay. *International Conference on Learning Representations (ICLR)*, 2016.

Juergen Schmidhuber. Artificial curiosity based on discovering novel algorithmic predictability through coevolution. In *Congress on Evolutionary Computation, IEEE Press*, 1999.

Wolfram Schultz, Paul Apicella, and Tomas Ljungberg. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. Journal of Neuroscience, 1993.

RJ Snowden, S Treue, RG Erickson, and RA Andersen. The response of area mt and v1 neurons to transparent motion. Journal of Neuroscience, 1991.

Richard Sutton. Temporal credit assignment in reinforcement learning. PhD Thesis University of Massachusetts Amherst, 1984.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. Cognitive Psychology, 1980.

Hado van Hasselt. Double q-learning. *Conference on Neural Information Processing Systems (NeurIPS)*, 2010.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando. De Freitas. Dueling network architectures for deep reinforcement learning. *Internation Conference on Machine Learning ICML.*, pp. 1995–2003, 2016.

Chris Watkins. Learning from delayed rewards. In *PhD thesis, Cambridge*. King's College, 1989.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. 1992.

Klaus Wunderlich, Antonio Rangel, and John P. O'Doherty. Neural computations underlying action-based decision making in the human brain. Proceedings of the National Academy of Sciences (PNAS) 2009.