Underlying Functional Structure of Reinforcement Learning

Ezgi Korkmaz

Abstract

Advances in deep reinforcement learning have enabled the development of policies capable of solving complex, high-dimensional problems, allowing AI agents to strategize and reason purely through interaction without supervision. Despite these successes, our knowledge on the functional properties and the underlying structure of the deep neural policy manifold remains limited. In this paper, we uncover a fundamental functional property in reinforcement learning: the intrinsic alignment between the advantage function and the gradient of the loss targeting directions of incoherence. We provide a rigorous theoretical foundation for this relationship, demonstrating that this intrinsic alignment characterizes how policies learn and internalize the underlying value function and how this process governs policy decisions. By leveraging this fundamental functional property, we propose a novel algorithm that can diagnose and identify reinforcement learning generalization, policy robustness and decision volatilities. We conduct extensive empirical analysis in high-dimensional MDPs. From algorithmic and architectural changes to natural distributional shifts and adversarial attacks, our proposed method can identify and audit the differences by leveraging the underlying structure of the deep neural policy manifold and the intrinsic alignment. Our analysis reveals foundational properties of policies trained in high-dimensional MDPs, and our paper provides a principled step toward constructing scalable, stable, and generalizable deep reinforcement learning agents with robust reasoning capabilities.

1 Introduction

Built on principled foundations [Watkins, 1989, Montague et al., 1996], reinforcement learning research has driven the creation of policies that can solve complex tasks purely via interaction in high-dimensional MDPs [Mnih et al., 2015, Korkmaz, 2025]. From highly complicated board games to mathematical and scientific reasoning of frontier large language models [Jaech et al., 2024, Guo et al., 2025], reinforcement learning has been the epicenter of agents that can strategize and surpass human cognitive skills. Nonetheless, while policy capabilities have advanced at an accelerated pace, our fundamental understanding of the functions learned by deep reinforcement learning policies and the underlying structure governing their behaviour and the internal dynamics of their learning processes remains remarkably limited.

This trajectory has resulted in the utilization of reinforcement learning to advance frontier large language models, marking a paradigm shift in AI that has enabled agents to move beyond complex strategizing toward the discovery of entirely novel scientific knowledge [Knuth, 2026, OpenAI, 2026a,b]. Yet, as we increasingly rely on these policies for high-stakes reasoning and decision-making, the *black-box* nature of their behaviour and their underlying reasoning creates a significant barrier to their full utilization, safety, and alignment. Specifically, we still lack principled explanations for why certain directions in the high-dimensional state space trigger sudden *volatilities* or reasoning failures. These instabilities suggest that while reinforcement learning successfully optimizes the

Correspondence to Ezgi Korkmaz: ezgikorkmazmail@gmail.com

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

manifold towards a policy with high performance, the underlying functional structure remains fragile, punctuated by regions where logical coherence abruptly breaks down. In this paper, we look beneath the surface of high performance to characterize the underlying properties of what is truly learned by deep reinforcement learning, its core principles, the structural drivers of policy failure, and make the following contributions:

Contributions. We first provide a theoretically well-founded rigorous analysis and characterize the underlying functional structure of reinforcement learning in Section 3. Our analysis reveals that there is an intrinsic alignment between the negative average of the advantage function and directions of incoherence in deep reinforcement learning. Our paper is the first one that demonstrates, both theoretically and empirically, the existence of this alignment, and further our analysis uncovers that this alignment is a direct consequence of the approximate convexity of the negative average advantage within the high-dimensional policy manifold. Our analysis reveals that the average negative advantage serves as a formal precursor to policy robustness, providing a principled characterization of deep neural policy manifold structure. Building on these insights, in Section 3, we introduce an algorithm that leverages this intrinsic alignment to identify and quantify uncertainties within an MDP. We then conduct extensive experiments in the Arcade Learning Environment in Section 4, and our results verify the theoretical analysis and further demonstrate that this alignment can be used to effectively diagnose the environment in which the policy interacts. Our work shows that, across a broad spectrum of challenges ranging from distributional shift to adversarial changes, this intrinsic alignment provides a high-fidelity signal for impending instabilities. Finally, we investigate different algorithms and architectural changes to examine how they influence the manifold's functional structure. Our analysis further demonstrates that algorithms that perform better have a tighter intrinsic alignment. Our results reveal a fundamental property of the deep reinforcement learning manifold and provide a principled approach to understand policy performance, stability and scalability.

2 Background and Preliminaries

Preliminaries. A Markov Decision Process is represented as a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \mu_0 \rangle$ where $\mathcal{S}$ represents the state space, $\mathcal{A}$ represents the action space, $r(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function which provides the rewards received when in state s the policy takes action a and transitions to state $\hat{s}$, γ represents the discount factor to prioritize short-term vs long-term rewards and $\mathcal{P}(s,a,\hat{s})$ on $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ is the transition probability function. The objective in reinforcement learning is to learn an optimal policy via interacting with an environment ε to maximize the expected cumulative rewards $\mathcal{R} = \mathbb{E}_{a_t \sim \pi(s_t,\cdot), s_{t+1} \sim \mathcal{P}(s_t,a_t,\cdot)} \sum_t \gamma^t r(s_t,a_t)$ obtained by the policy $\pi(s,a)$. $\mathcal{Q}$-learning achieves this objective via iterative Bellman update [Bellman, 1957]. Precisely, $\mathcal{Q}$-function is updated via

$$\mathcal{B}^\pi \mathcal{Q}(s,a) = \mathbb{E}[r(s,a)] + \gamma \mathbb{E}_{a \sim \pi(s,\cdot), \hat{s} \sim \mathcal{P}(s,a,\cdot)} \max_{\hat{a}} \mathcal{Q}(\hat{s},\hat{a}).$$

$\mathcal{B}^\pi$ represents the *Bellman operator* and $\mathcal{V}(s) = \max_a \mathcal{Q}(s,a)$ determines the value of the state.

Computation of Adversarial Directions. The vulnerabilities of deep neural networks have been demonstrated in the early work of Szegedy et al. [2014]. By adding imperceptible perturbations computed by a box-constrained optimization to the sample, the authors were able to change the decision of the model. While this method is computationally draining, later the work of Goodfellow et al. [2015] proposed the fast gradient sign method via linearizing the cost function at the sample where the adversarial direction needs to be computed. Further, this objective has been extended to an iterative version by Kurakin et al. [2016].

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

Another natural extension of optimizing this objective was first extended to momentum iterative [Dong et al., 2018], and later is computed via Nesterov Momentum [Korkmaz and Brown-Cohen, 2023]

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

where μ is the decaying factor. Another line of adversarial direction computation is focused on computing the minimum distance to a decision boundary where the optimal decision of the policy is not executed under the adversarial direction. In particular, Carlini and Wagner [2017] formulation targets to optimize the objective $\min_{x_{\text{adv}} \in \mathcal{X}} c \cdot J(x_{\text{adv}}) + \|x - x_{\text{adv}}\|_2^2$ The elastic-net regularization (i.e. ENR) is the same objective of the Carlini and Wagner [2017] formulation with the ℓ_1-norm regularized version [Chen et al., 2018].

$$\min_{x_{\text{adv}} \in \mathcal{X}} c \cdot J(x_{\text{adv}}) + \kappa_1 \|x - x_{\text{adv}}\|_1 + \kappa_2 \|x - x_{\text{adv}}\|_2^2$$

Issues with Robustness and Functional Instabilities of Deep Reinforcement Learning. Early studies in the adversarial view demonstrated that deep neural policies are vulnerable to imperceptible perturbations introduced to their input [Huang et al., 2017] computed via the fast gradient sign method [Goodfellow et al., 2015]. Subsequent studies expanded on this by employing sophisticated optimization techniques, such as Nesterov Momentum. Following this line of work, several studies demonstrated via the Carlini and Wagner [2017] and ENR formulation [Chen et al., 2018] that deep reinforcement learning policies learn adversarial features that generalize across different Markov Decision Processes (MDPs) and algorithmic architectures including those specifically trained for certified robustness [Korkmaz, 2022]. To address these volatility issues of reinforcement learning a large body of work focused on techniques to obtain robustness. In particular, Pinto et al. [2017], Gleave et al. [2020] proposed to model this interaction as zero-sum Markov game and jointly train the policy in the presence of an adversary. However, recent work has begun to expose the hidden costs of these defenses. Adversarial training is known to suffer from significant drawbacks, ranging from poor generalization to vulnerability against black-box attacks. Specifically, research indicates that robustly trained deep reinforcement learning policies fail to match the generalization capabilities of their vanilla-trained counterparts [Korkmaz, 2023, 2024]. Furthermore, recent evidence demonstrates that adversarial training induces reasoning and counterfactual errors in robust reinforcement learning that are notably absent in standard reinforcement learning [Korkmaz, 2026]. Despite a growing catalog of empirical failures ranging from generalization to reasoning and counterfactuality, current paradigms remain limited to documenting these observations without providing a foundational analysis of the precise mechanisms and underlying structures that cause reinforcement learning to experience such failure modes.

3 The Intrinsic Alignment of Average Negative Advantage and Deep Neural Policy Manifold Curvature

Analyzing the functional properties and behavioral dynamics of reinforcement learning has long been at the center of discussion [Schmidhuber, 1991, Tesauro, 1991, Baird, 1995, Singh and Dayan, 1996]. To truly understand the underlying structure of reinforcement learning, in this section we will provide a theoretical analysis of the functional properties of the deep neural policy manifold. The primary quantity in our analysis is the canonical loss $\mathcal{J}$ used to identify directions of incoherence on the deep neural policy manifold. Given a state-action value function $\mathcal{Q}(s,a)$ in state s, the argmax policy takes the action $a^*(s) = \arg\max_{a \in \mathcal{A}} \mathcal{Q}(s,a)$. The softmax policy $\pi_{\mathcal{Q}}(s,a)$ is given by the softmax function of the $\mathcal{Q}$-values for each action in state s i.e. $\pi_{\mathcal{Q}}(s,a) = \frac{\exp \mathcal{Q}(s,a)}{\sum_{a' \in \mathcal{A}} \exp \mathcal{Q}(s,a')}$.

Definition 3.1 (*Policy Coherence Loss*)**.** The instability loss between states s and $\hat{s}$ is given by the cross-entropy between the argmax policy in state $\hat{s}$ and the softmax policy in state s,

$$\mathcal{J}(s,\hat{s}) = -\log \pi_{\mathcal{Q}}(s, \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a))$$

Intuitively $\mathcal{J}(s,\hat{s})$ is a smooth measure of how far the policy in state s has deviated from the argmax policy in state $\hat{s}$. Observe that since $\mathcal{Q}(s,a)$ is differentiable with respect to s then so is $\mathcal{J}$. Thus, the gradient $\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}}$ corresponds to the direction along which the policy most rapidly deviates from the argmax policy in state $\hat{s}$. The second quantity we study is the gap between the maximum $\mathcal{Q}$-value and the average $\mathcal{Q}$-value in each state s.

Definition 3.2 (*Average Advantage Value*)**.** The average advantage value, denoted as $\Omega(s,\hat{s})$, is the difference between the $\mathcal{Q}$-value of the optimal action from state $\hat{s}$ evaluated at state s and the expected $\mathcal{Q}$-value over the action space $\mathcal{A}$ at state s:

$$\Omega(s,\hat{s}) = \mathcal{Q}(s, \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a)) - \mathbb{E}_{a \sim \mathcal{A}} \mathcal{Q}(s,a)$$

Thus, the gap between the $\mathcal{Q}$-value of the optimal action in state s and the average action in state s is given by the evaluation of $\Omega(s,\hat{s})$ with $\hat{s} = s$ (i.e. $\Omega(s,s)$). In this section, we will analyze the underlying structure and functional properties of reinforcement learning, and reveal the intrinsic alignment between average advantage value, i.e. $\Omega(s,s)$, and directions of incoherence, i.e. the norm of the gradient of the policy coherence loss $\|\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}}\|$.

Proposition 3.3 (*Directions of Incoherence*)**.** *Let $\pi_{\mathcal{Q}}(s,a) = \frac{\exp \mathcal{Q}(s,a)}{\sum_{a' \in \mathcal{A}} \exp \mathcal{Q}(s,a')}$ be the softmax policy defined by $\mathcal{Q}$. Let $a^*(\hat{s}) = \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a)$. Then the gradient of the coherence loss $\mathcal{J}(s,\hat{s})$ is given by*

$$\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} = \mathbb{E}_{a \sim \pi_{\mathcal{Q}}(\hat{s},a)}[\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}] - \nabla_s \mathcal{Q}(s, a^*(\hat{s}))|_{s=\hat{s}}$$

Hence, the proximity of the directions of the incoherence on the deep neural policy manifold $-\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}}$ to the gradient of the advantage function $\nabla_s \Omega(s,\hat{s})|_{s=\hat{s}}$ provides the connection between $-\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}}$, $\nabla_s \Omega(s,\hat{s})|_{s=\hat{s}}$ and the softmax policy $\pi_{\mathcal{Q}}(s,a)$ (see proof in A.1).

Lemma 3.4 (*Proximity of Gradient of Average Advantage to Directions of Incoherence*)**.** *Let $\epsilon > 0$ and suppose that $\sum_{a\in\mathcal{A}} |\pi_{\mathcal{Q}}(\hat{s},a) - \frac{1}{|\mathcal{A}|}| < \epsilon$. Then*

$$\|-\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} - \nabla_s \Omega(s,\hat{s})|_{s=\hat{s}}\| < \epsilon \max_{a\in\mathcal{A}} \|\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}\|.$$

Proof. By Proposition 3.3,

$$\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} = \mathbb{E}_{a\sim\pi_{\mathcal{Q}}(\hat{s},a)}[\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}] - \nabla_s \mathcal{Q}(s, \arg\max_{a\in\mathcal{A}} \mathcal{Q}(\hat{s},a))|_{s=\hat{s}}.$$

Next, the gradient of Ω is given by

$$\nabla_s \Omega(s,\hat{s})|_{s=\hat{s}} = \nabla_s \mathcal{Q}(s, \arg\max_{a\in\mathcal{A}} \mathcal{Q}(\hat{s},a))|_{s=\hat{s}} - \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}$$

Combining the above two formulas yields

$$\begin{aligned}
&\|-\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} - \nabla_s \Omega(s,\hat{s})|_{s=\hat{s}}\| = \left\|\sum_{a\in\mathcal{A}} (\pi_{\mathcal{Q}}(\hat{s},a) - \frac{1}{|\mathcal{A}|})\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}\right\| \\
&\leq \sum_{a\in\mathcal{A}} \left\|(\pi_{\mathcal{Q}}(\hat{s},a) - \frac{1}{|\mathcal{A}|})\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}\right\| \leq \sum_{a\in\mathcal{A}} |\pi_{\mathcal{Q}}(\hat{s},a) - \frac{1}{|\mathcal{A}|}| \cdot \max_{a\in\mathcal{A}} \|\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}\| \\
&\leq \epsilon \max_{a\in\mathcal{A}} \|\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}\|
\end{aligned}$$

□

Thus, Lemma 3.4 shows that a correlation between $\|\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}}\|$ and $\Omega(s,s)$ is in fact a correlation between $\|\nabla_s \Omega(s,\hat{s})|_{s=\hat{s}}\|$ and $\Omega(s,s)$. Next we show that this alignment can be entirely explained by $\Omega(s,\hat{s})$ being a convex function of s. In particular, along every curve of steepest ascent for a strictly convex function Ψ, the value of $\|\nabla\Psi\|$ is increasing.

Theorem 3.5 (*Gradient Alignment from Convexity*)**.** *Let $\Psi : \mathbb{R}^n \to \mathbb{R}$ be strictly convex with Lipschitz continuous gradients, $s_0 \in \mathbb{R}^n$, and $\zeta > 0$. Let the curve $s : [0,\zeta] \to \mathbb{R}^n$ be the (unique) solution of $s'(\xi) = \nabla\Psi(s(\xi))$ and $s(0) = s_0$. Then for all $\xi_1, \xi_2 \in [0,\zeta]$ with $\xi_1 < \xi_2$ we have $\|\nabla\Psi(s(\xi_1))\| < \|\nabla\Psi(s(\xi_2))\|$.*

Proof. Note that Lipschitz continuity of $\nabla\Psi$ implies that the differential equation with boundary data $s(0) = s_0$ defining $s(\xi)$ has a unique solution. The multivariate chain rule implies that

$$\frac{d}{d\xi}\|\nabla\Psi(s(\xi))\|^2 = 2\nabla\Psi(s(\xi))^\top \left(\nabla^2\Psi(s(\xi))\right) s'(\xi). \tag{1}$$

Applying the fundamental theorem of calculus, followed by (1) yields

$$\begin{aligned}
\|\nabla\Psi(s(\xi_2))\|^2 - \|\nabla\Psi(s(\xi_1))\|^2 &= \int_{\xi_1}^{\xi_2} \frac{d}{d\xi}\|\nabla\Psi(s(\xi))\|^2 d\xi = \int_{\xi_1}^{\xi_2} 2\cdot\nabla\Psi(s(\xi))^\top \left(\nabla^2\Psi(s(\xi))\right) s'(\xi) d\xi \\
&= \int_{\xi_1}^{\xi_2} 2\cdot\nabla\Psi(s(\xi))^\top \left(\nabla^2\Psi(s(\xi))\right) \nabla\Psi(s(\xi)) d\xi
\end{aligned}$$

where the final equality used the fact that $s(\xi)$ satisfies $s'(\xi) = \nabla\Psi(s(\xi))$. Next, the strict convexity of Ψ implies that $\nabla^2\Psi$ is positive definite. Hence, $2\cdot\nabla\Psi(s(\xi))^\top \left(\nabla^2\Psi(s(\xi))\right)\nabla\Psi(s(\xi)) > 0$. Thus we conclude that

$$\|\nabla\Psi(s(\xi_2))\|^2 - \|\nabla\Psi(s(\xi_1))\|^2 = \int_{\xi_1}^{\xi_2} 2\cdot\nabla\Psi(s(\xi))^\top \left(\nabla^2\Psi(s(\xi))\right) \nabla\Psi(s(\xi)) d\xi > 0$$

which implies that $\|\nabla\Psi(s(\xi_1))\| < \|\nabla\Psi(s(\xi_2))\|$ as desired. □

Deep reinforcement learning in high-dimensional MDPs learns approximately convex policies.

Theorem 3.5 shows that the observed intrinsic alignment is a direct consequence of the convexity of $\Omega(s,\hat{s})$. At first this result might seem quite surprising and counterintuitive; however, a closer examination of the training dynamics and the underlying design paradigms of deep reinforcement learning algorithms and their optimization reveals a clear underlying rationale: starting from utilization of ReLUs [Glorot et al., 2011] to network preferences [Goodfellow et al., 2012] all are specifically tuned to operate predominantly within their non-saturating regimes. Theorem 3.5 shows that along every

curve of steepest ascent for $\Omega(s, s)$, the norm of the gradient of $\Omega(s, s)$ will increase. To gain further intuition on the underlying functional structure of the average advantage with respect to s, consider two states $s, \tilde{s} \in \mathcal{S}$ with the same optimal action $a^* = \arg\max_{a \in \mathcal{A}} \mathcal{Q}(s, a) = \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\tilde{s}, a)$. For $\mathcal{Z} \in [0, 1]$, let $s_{\mathcal{Z}} = (1 - \mathcal{Z})s + \mathcal{Z}\tilde{s}$ be the linear interpolation between s and $\tilde{s}$. While $s, \tilde{s}$ are valid state observations of the MDP, the intermediate interpolated states $s_{\mathcal{Z}}$ are instead an unnatural, physically unrealizable combination of two valid states. The $\mathcal{Q}$-function has been trained to have a gap between the value of a^* and the non-optimal actions at the endpoints s and $\tilde{s}$, this functional separation degrades in the unnatural interpolated regions. Consequently, the average gap $\Omega(s_{\mathcal{Z}}, s)$ tends to dip as $\mathcal{Z}$ moves away from 0, only recovering as it approaches to $\mathcal{Z} = 1$ where it reaches $\Omega(\tilde{s}, s)$. That is, along the interpolation between s and $\tilde{s}$, the value of $\Omega(s_{\mathcal{Z}}, s)$ lies below the straight line connecting $\Omega(s, s)$ and $\Omega(\tilde{s}, s)$ i.e. $\Omega((1 - \mathcal{Z})s + \mathcal{Z}\tilde{s}, s) \leq (1 - \mathcal{Z})\Omega(s, s) + \mathcal{Z}\Omega(\tilde{s}, s)$. That is, Ω is convex on the line between s and $\tilde{s}$. More generally, given any set of k states $s_1, s_2, \ldots s_k$ in $n > k$ dimensions, this property of Ω holds over the simplex defined by the k states, because any convex combination of the s_i will on average cause the $\mathcal{Q}$-function to be less decisive and hence have a lower value for the average gap Ω. In this context, Section 4.3 and Section 4.4 detail how the intrinsic alignment of average advantage and directions of incoherence decreases in unnatural states, spanning the spectrum from adversarial perturbations to distributional shifts. Now we will prove a stronger result on the functional property of Ω in high-dimensional state spaces.

Theorem 3.6 (Strict Curvature Extension [Yan, 2014])**.** *Let X be a compact convex subset of $\mathbb{R}^n$, and $F : \mathbb{R}^n \to \mathbb{R}$ a twice continuously differentiable function with positive definite Hessian on X. Then F can be extended to a twice continuously differentiable function with positive definite Hessian on all of $\mathbb{R}^n$.*

Leveraging this result, we show that convexity of Ω on a simplex of states, which as described above is a result of training, implies that Ω can be extended to a convex function on the entire state space.

Theorem 3.7 (*Strict Convexity of Negative Average Advantage*)**.** *Let S^π be a finite sequence of states visited by the greedy policy π with respect to state-action value function $\mathcal{Q}$ in an MDP M. Assume that the state space $\mathcal{S}$ has dimension $n \geq |S^\pi|$, and that $\Omega(s, \hat{s})$ is a twice continuously differentiable function of s and strictly convex on the simplex $\mathrm{conv}(S^\pi)$. Then Ω can be extended to a strictly convex function on the entire state space.*

The proof of Theorem 3.7 is provided in the Appendix B. Intuitively, Theorem 3.7 states that if we take a trajectory of states in a high-dimensional state-space MDP, then assuming that Ω is strictly convex on the convex hull of the trajectory of states implies that Ω can be extended to a convex function on the whole state space. This functional property of Ω likely arises as a byproduct of training, where actual valid state observations show a clear gap in the advantage of a^* over the average action, while intermediate interpolated states are likely to lack this property. Therefore, by Theorem 3.7 the values of Ω at the observed states in the trajectory are equal to the values of a strictly convex function which extends Ω to the entire state-space.

Theorem 3.5 predicts $\Omega(s, s)$ and $\|\nabla_s \Omega(s, \hat{s})|_{s=\hat{s}}\|$ to be positively aligned. These theoretical predictions are subsequently verified in Section 4 in the experimental analysis. We further propose an algorithm described in Algorithm 1 to identify and diagnose the environment with which the agent interacts by leveraging this intrinsic alignment. We further investigate how we can identify adversarial disturbances, distributional shifts and even algorithmic differences via the intrinsic alignment in Section 4.3 and Section 4.4.

Algorithm 1 SCALE: Structural Coherence & Alignment Evaluation in Deep Reinforcement Learning

Input: State action value function $\mathcal{Q}(s, a)$, expected value μ and standard deviation σ of Spearman correlation from unperturbed trajectory, and tolerance k.
for $t = 1$ **to** T **do**
  Set $\mathcal{G}_t = \|\nabla_s \mathcal{J}(s, s_t)|_{s=s_t}\|$ and $\Omega_t = \Omega(s_t, s_t)$
  Compute next action $a_t = \arg\max_{a \in \mathcal{A}} \mathcal{Q}(s_t, a)$.
  Observe next state $s_{t+1} \sim \mathcal{P}(s_t, a_t, \cdot)$
end for
Compute ρ_{spearman} of $\{(\Omega_t, \mathcal{G}_t)\}_{t=1}^T$.
Return: "Perturbation present" if $|\rho_{\text{spearman}} - \mu| > k \cdot \sigma$.

4 Experimental Analysis

The empirical analysis in this section verifies the theoretical analysis provided in Section 3. First, the results demonstrate that there is an alignment between $\Omega(s, \hat{s})$ and $\nabla_s \mathcal{J}(s, \hat{s})$, lending support to the claim regarding the underlying functional structure of $\Omega(s, \hat{s})$. Second, for trajectories of states that have been perturbed, either adversarially or by semantically meaningful perturbations to observations, the positive correlation is lower. This further confirms the theoretical arguments in Section 3, where

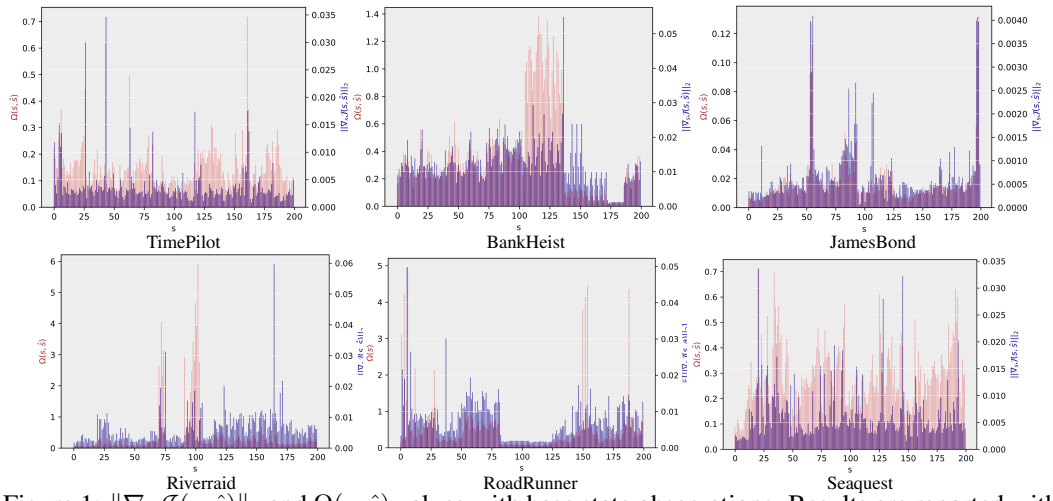


Figure 1: $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$ and $\Omega(s, \hat{s})$ values with base state observations. Results are reported with double y-axis bar graph where each axis color matches its corresponding bars.

Reinforcement learning has an intrinsic alignment between the directions of incoherence of the deep neural policy manifold and the negative average advantage.

it is predicted that the intrinsic alignment of Ω should be stronger at valid states that are similar to those seen in the training distribution than at perturbed states. The deep reinforcement learning policies are trained via double $\mathcal{Q}$-learning proposed in van Hasselt [2010] and later implemented in van Hasselt et al. [2016] with dueling architecture [Wang et al., 2016] in high-dimensional MDPs. For the results reported in Section 4.1 several policies additionally are trained with the prioritized algorithm [Schaul et al., 2016]. Adversarial directions are computed via the C&W formulation, ENR and Nesterov momentum. The hyperparameters for C&W are learning rate 0.01, maximum iteration 1000, initial constant 10 and for Nesterov momentum iteration is 100. Note that we used the exact hyperparameter settings as prior work for the adversarial directions to be able to provide consistent and transparent results. Our empirical results will use two standard metrics to measure non-linear correlations between variables: Spearman and Kendall's tau. Spearman correlation for a sample of n pairs $\{(X_i, Y_i)\}_{i=1}^n$ is computed by first calculating the ranks $R(X_i)$ and $R(Y_i)$, where R assigns the numerical rank $\{1, 2, \dots n\}$ of the ranking of the X_i variables from smallest to largest (and similarly for the Y_i variables). Then the Spearman correlation is given by calculating the standard Pearson correlation of the rank variables $\rho_{R(X),R(Y)} = \frac{\text{Cov}(R(X),R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$. Hence a perfectly monotone function $Y = f(X)$ will have Spearman correlation equal to 1.0. To define Kendall's tau let C denote the number of pairs of samples (X_i, Y_i), (X_j, Y_j) where either $X_i > X_j$ and $Y_i > Y_j$ or $X_i < X_j$ and $Y_i < Y_j$. Let D denote the number of remaining pairs, that do not satisfy the above property. Kendall's tau is given by $\tau(X, Y) = \frac{C-D}{n(n-1)/2}$. In particular, the numerator is the number of pairs where both coordinates are consistently ordered, minus the number of pairs inconsistently ordered, and the denominator is the total number of pairs. As with Spearman correlation, a perfectly monotone function $f(X, Y)$ will have $\tau(X, Y) = 1$. Table 1 reports the Spearman, Pearson and Kendall's τ results of the intrinsic correlation. These results verify the theoretical analysis provided in Section 3. Beyond the analytical probing of the deep neural policy landscape, a more intuitive understanding of neural processing can be gained by examining natural intelligence cognitive responses to visual illusions. This approach serves as a canonical approach in neuroscientific research [Hubel and Wiesel, 1962, Grunewald and Lankheet, 1996, Westheimer, 2008], enabling the anatomical diagnosis of processing irregularities to specific brain regions, from parahippocampal cortex to cortical areas [Axelrod et al., 2017, Seymour et al., 2018]. Our approach,

Table 1: Spearman, Kendall's τ and Pearson correlations between $\Omega(s, \hat{s})$ and $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$ with base state observations.

MDPs	Base Spearman	Base Kendall's τ	Base Pearson
JamesBond	0.80887±0.01370	0.61751±0.01242	0.73822±0.01812
Riverraid	0.75758±0.00544	0.55706±0.00348	0.19795±0.01407
RoadRunner	0.66325±0.01640	0.46928±0.01421	0.40010±0.00854
BankHeist	0.81963±0.01104	0.62919±0.00990	0.53146±0.00047
Seaquest	0.63826±0.00540	0.45793±0.00505	0.56361±0.02136
TimePilot	0.94431±0.00357	0.81207±0.00830	0.86280±0.00973

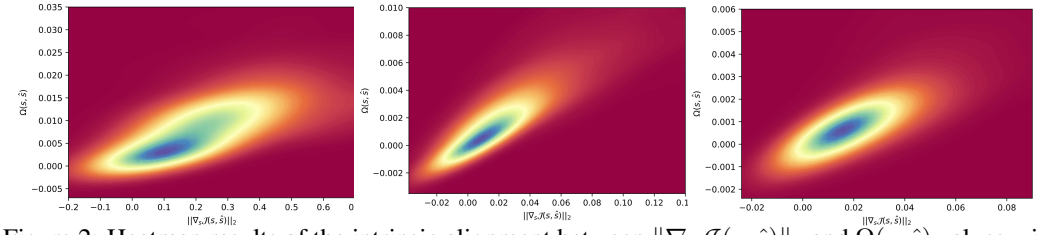


Figure 2: Heatmap results of the intrinsic alignment between $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$ and $\Omega(s, \hat{s})$ values with base state observations in TimePilot, JamesBond and BankHeist.

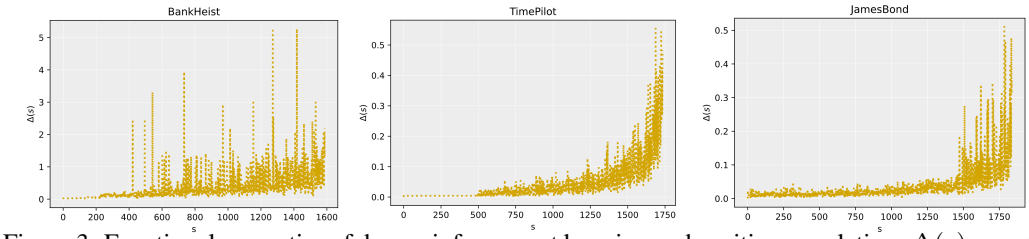


Figure 3: Functional properties of deep reinforcement learning and positive correlation: $\Delta(s)$ reported with base state observations in high-dimensional state representation MDPs.

Table 2: Architectural differences and the structural properties of deep reinforcement learning: the Spearman, Pearson and Kendall's τ correlations between $\Omega(s, \hat{s})$ and $\mathbb{E}[\|\nabla_s \mathcal{J}(s, \hat{s})\|_2]$.

Environment	RoadRunner		BankHeist	
Architecture	Prior	Duel	Prior	Duel
Spearman	0.57537±0.048449	**0.66325±0.016404**	0.61924±0.011326	**0.81963±0.0110428**
Kendall's τ	0.40126±0.037366	**0.46928±0.014211**	0.42684±0.008932	**0.62919±0.0099065**
Pearson	0.25306±0.036695	**0.40010±0.008546**	0.42745±0.022079	**0.53146±0.0004740**

predicated on analyzing directions of incoherence, draws a conceptual parallel to the neurobiological processing of visual illusionary stimuli and our results reveal the capabilities of our method on not only the identification of natural or adversarial changes but further diagnosis of algorithmic and architectural differences.

4.1 Architectural and Algorithmic Differences

In this section, we examine how different algorithmic designs and architectural choices influence the functional structure of the policy manifold and specifically the intrinsic alignment between directions of incoherence and the average advantage. In particular, Table 2 reports the correlation results with architectural differences between dueling and prior architecture, and the results demonstrate that the dueling architecture has a tighter alignment. This performance gap reveals a **fundamental structural correspondence**: the dueling architecture which achieves a human-normalized median score of **172.1%** compared to 123.7% for the prior architecture [Wang et al., 2016], exhibits a significantly more rigid intrinsic alignment. Thus, these results further solidify that the algorithms that perform better have a tighter alignment between $\Omega(s, \hat{s})$ and $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$. The results reported in Table 2 demonstrate that SCALE effectively identifies how architectural and algorithmic variations change underlying functional structure and the intrinsic alignment on the deep neural policy manifold. Specifically, our analysis reveals that the approximate convexity of the advantage function is significantly more pronounced in higher-performing algorithms. Consequently, these findings establish that the degree of the intrinsic alignment is a fundamental property that can be targeted through architectural or algorithmic refinement to systematically enhance reinforcement learning performance.

4.2 The Underlying Structure of the Deep Reinforcement Learning Manifold

In this section, our objective is to leverage the functional properties of the deep neural policy manifold to diagnose and understand the reinforcement learning policy coherence and decision making. In particular, Figure 1 reports $\Omega(s, \hat{s})$ and $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$ values for Riverraid, RoadRunner, Seaquest, TimePilot, JamesBond and BankHeist with base state observations. While these results provide a qualitative alignment between the average advantage and the incoherence gradient norm $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$, we provide a more rigorous quantitative assessment in Table 1, reporting Spearman,

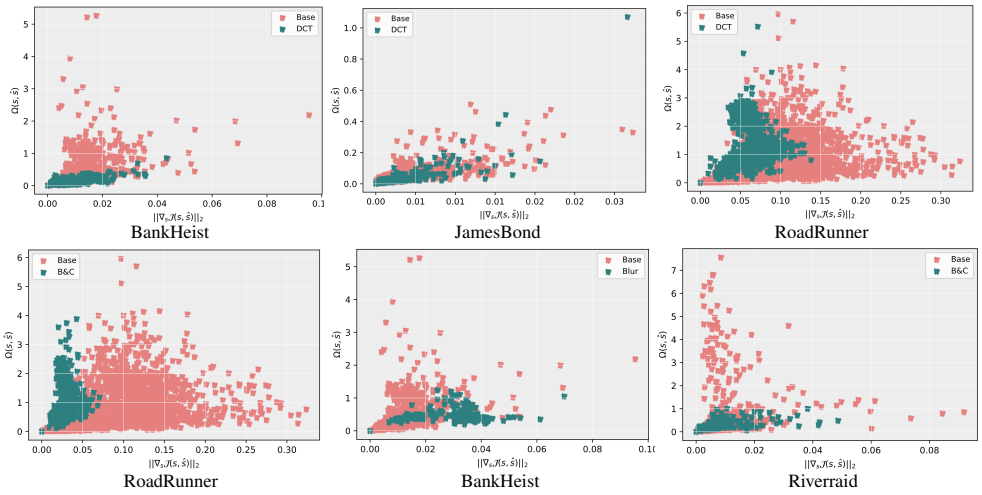


Figure 4: Scatter plots of $\Omega(s, \hat{s})$ and $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$ with base states and distributional shift, i.e. policy-independent imperceptible natural perturbations, computed via B&C, Blur and DCT.

Table 3: Spearman and Kendall's τ correlations of $\Omega(s, \hat{s})$ and $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$ with base observations and in the existence of adversarial directions computed via C&W, ENR and Nesterov Momentum.

Correlations	JamesBond	Riverraid	RoadRunner
Base Spearman	**0.8088733±0.013700**	**0.7575881±0.005440**	**0.663257±0.016404**
ElasticNet (ENR) Spearman	0.7404537±0.007845	0.7312598±0.007944	0.360172±0.0161319
Nesterov Momentum Spearman	0.737131±0.005016	0.618266±0.0332335	0.507291±0.0103992
Carlini and Wagner Spearman	0.7774369±0.0034135	0.5979524±0.019333	0.359959±0.0443513
Base Kendall's τ	**0.61751±0.012425**	**0.557067±0.003489**	**0.469280±0.014211**
ElasticNet (ENR) Kendall's τ	0.5518718±0.007701	0.5276511±0.006757	0.246347±0.010676
Nesterov Momentum Kendall's τ	0.548718±0.0055495	0.509211±0.028019	0.325897±0.020279
Carlini and Wagner Kendall's τ	0.5897872±0.005087	0.4256483±0.014264	0.244640±0.030973
Correlations	BankHeist	Seaquest	TimePilot
Base Spearman	**0.8196393±0.0110428**	**0.6382674±0.0054034**	**0.9443157±0.003573**
ElasticNet (ENR) Spearman	0.6285715±0.02395142	0.5775322±0.0091588	0.9294928±0.008797
Nesterov Momentum Spearman	0.6565451±0.03250028	0.5542451±0.0126013	0.9060342±0.0140901
Carlini and Wagner Spearman	0.373751±0.0038957	0.4573776±0.0092046	0.7290117±0.006987
Base Kendall's τ	**0.629199±0.0099065**	**0.4579388±0.005053**	**0.8120791±0.0083035**
ElasticNet (ENR) Kendall's τ	0.4532070±0.01841255	0.4118854±0.0063908	0.7887973±0.0126924
Nesterov Momentum Kendall's τ	0.4642903±0.03478681	0.382152±0.0087899	0.7507642±0.0185963
Carlini and Wagner Kendall's τ	0.2994147±0.0393197	0.3155310±0.0054738	0.5567044±0.006622

Pearson, and Kendall's τ correlation coefficients. Furthermore, Figure 2 visualizes the kernel density estimation of points with xy-coordinates given by $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$ and $\Omega(s, \hat{s})$, providing an alternative visualization of the alignment between these two quantities. To give an alternative view on the correlation, we sort the states s in increasing order by the magnitude of $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$. Let $\Delta(s)$ for $s = 1, 2, \ldots$ to indicate the value of $\Omega(s, s)$ on the 1st,2nd,... state in this sorted order. Figure 3 further reports the values of $\Delta(s)$ as s ranges over a trajectory of states sorted by increasing $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$, providing clear additional evidence of a positive correlation.

4.3 The Intrinsic Alignment and Adversarial Perturbations

In this section, we investigate the underlying structure of the deep neural policy manifold when the deep reinforcement learning policy is exposed to adversarial perturbations. In particular, Table 3 gives the Spearman correlation and Kendall's tau results for adversarially perturbed states alongside base, i.e. unperturbed, states. These results once more verify the theoretical predictions in Section 3, where the intrinsic alignment between average advantage and the directions of incoherence is tighter at the base state observations compared to perturbed state observations. The clear disparities in alignment between unperturbed trajectories and those subjected to adversarial stress enable the robust detection of environmental shifts. Specifically, Algorithm 1 facilitates the real-time detection of such perturbations, while the results in Table 3 demonstrate that SCALE not only identifies the existence of adversarial perturbations, but also distinguishes between specific attack methodologies based on

Table 4: Spearman and Kendall's τ correlations between $\Omega(s, \hat{s})$ and $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$ with base state observations and in the existence of policy-independent (i.e. semantically meaningful) non-robust directions computed via DCT, B&C, Blur.

Correlations	JamesBond	Riverraid	RoadRunner
Base Spearman	**0.808873±0.013700**	**0.757588±0.005440**	**0.66325±0.016404**
Blur Spearman	0.739044±0.006811	0.677357±0.017423	0.498806±0.039418
B&C Spearman	0.711883±0.007793	0.670667±0.023612	0.384988±0.024496
DCT Spearman	0.595459±0.066479	0.639586±0.012965	0.322124±0.104422
Base Kendall's τ	**0.6175157±0.012425**	**0.557067±0.003489**	**0.469280±0.014211**
Blur Kendall's τ	0.545957±0.007139	0.507149±0.027933	0.340502±0.024998
B&C Kendall's τ	0.522245±0.0071492	0.493685±0.024185	0.261689±0.018721
DCT Kendall's τ	0.477147±0.0433372	0.454356±0.010088	0.218891±0.066816
Correlations	BankHeist	Seaquest	TimePilot
Base Spearman	**0.819639±0.0110428**	**0.6382674±0.0054034**	**0.9443157±0.003573**
Blur Spearman	0.2226758±0.1330715	0.5311349±0.0145224	0.9086324±0.015733
B&C Spearman	0.5663207±0.0506708	0.2993260±0.0580819	0.5897620±0.018235
DCT Spearman	0.0365417±0.1172801	0.4326550±0.0230028	0.9000135±0.008147
Base Kendall's τ	**0.629199±0.0099065**	**0.4579388±0.005053**	**0.8120791±0.0083035**
Blur Kendall's τ	0.1620834±0.0896341	0.3702977±0.0108535	0.7153446±0.032583
B&C Kendall's τ	0.4062786±0.0468716	0.2009678±0.0394226	0.4020167±0.004175
DCT Kendall's τ	0.0216564±0.0869163	0.2993766±0.0168171	0.7524561±0.012622

their unique structural signatures. For instance, SCALE precisely differentiates the impact of the C&W attack on intrinsic manifold correlations compared to the EAD attack, providing a granular diagnostic of policy degradation.

4.4 Distributional Shift: Imperceptible Natural Non-Robust Perturbations

In this section, we further investigate how correlations change under imperceptible distributional shift, i.e. natural non-robust directions. In particular, Table 4 reports Spearman and Kendall's τ correlation results under policy-independent non-robust directions computed via DCT, B&C and blur. The policy-independent non-robust directions are computed within perceptual similarity bound to ensure that these perturbations are as invisible as adversarial directions to the human perception. These imperceptible distributional shifts recently have been shown to break the certified robust defences in deep reinforcement learning [Korkmaz, 2023]. In particular, the perceptual similarity $\Phi_{\text{similarity}}(s, \hat{s})$ is computed via $\Phi_{\text{similarity}}(s, \hat{s}) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}^l_{shw} - \hat{y}^l_{\xi(s)hw})\|_2^2$ where W_l, H_l, and C_l is the width, height and number of channels in the convolutional layers respectively and $\hat{y}^l_s, \hat{y}^l_{\Psi(s)} \in \mathbb{R}^{W_l \times H_l \times C_l}$ is the vector of the unit normalized activations. Figure 4 reports the values of $\Omega(s, \hat{s})$ and $\|\nabla_s \mathcal{J}(s, \hat{s})\|_2$ under different policy-independent non-robust directions. These results once more demonstrate that from adversarial attacks to distributional shift our proposed algorithm SCALE can identify and diagnose the environment the reinforcement learning policy experiences, thereby allowing the policy to act according to the risks and instabilities currently present in the environment.

5 Conclusion

In this paper we focus on the underlying structure of the deep neural policy manifold and inherent functional properties of deep reinforcement learning. Our paper discovers that there is an intrinsic correlation between negative average advantage and the directions of incoherence. We provide a rigorous foundational analysis and establish that this alignment is a direct consequence of the approximate convexity of the deep reinforcement learning policy manifold. We conduct extensive experiments in the Arcade Learning Environment with high-dimensional state observations. Our empirical results lay out how this intrinsic correlation changes across a wide range of adversarial attacks and natural non-robust directions, i.e. distributional shift, and further across different training algorithms and architectural changes. Our paper further introduces an algorithm that leverages this intrinsic alignment to assess policy generalization and robustness, diagnose the environment the deep neural policy is interacting and its effects on the policy decision making. The theoretical and empirical analysis provided in our work lays out the intrinsic properties of the deep reinforcement

learning policy landscape that can be directly leveraged to obtain scalable, stable and robust policies that can generalize and reason.

References

V. Axelrod, D. S. Schwarzkopf, S. Gilaie-Dotan, and G. Rees. Perceptual similarity and the neural correlates of geometrical illusions in human brain structure. *Nature Scientific Reports*, 2017.

L. Baird. Residual algorithms: Reinforcement learning with function approximation. *Machine Learning Proceedings*, 1995.

R. E. Bellman. Dynamic programming. In *Princeton, NJ: Princeton University Press*, 1957.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *In 2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

P. Chen, Y. Sharma, H. Zhang, J. Yi, and C. Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. In S. A. McIlraith and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 10–17. AAAI Press, 2018.

Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

European Union Parliament. Eu ai act: first regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, 2023.

Europen Union. United states approach to artificial intelligence. 2024.

A. Gleave, M. Dennis, C. Wild, K. Neel, S. Levine, and S. Russell. Adversarial policies: Attacking deep reinforcement learning. *International Conference on Learning Representations ICLR*, 2020.

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In G. J. Gordon, D. B. Dunson, and M. Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pages 315–323. JMLR.org, 2011.

I. Goodfellow, J. Shelens, and C. Szegedy. Explaning and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2012, Atlanta, GA, USA, 16-21 June 2012*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1319–1327. JMLR.org, 2012.

A. Grunewald and M. J. M. Lankheet. Orthogonal motion after-effect illusion predicted by a model of cortical motion processing. *Nature*, 1996.

D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Ding, H. Gao, M. Zhang, M. Zhang, M. Tang, M. Zhou, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan,

Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

S. Huang, N. Papernot, Y. Goodfellow, Ian an Duan, and P. Abbeel. Adversarial attacks on neural network policies. *Workshop Track of the 5th International Conference on Learning Representations*, 2017.

D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 1962.

A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A. T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Zhang, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C. M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F. P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H. W. Chung, I. Kivlichan, I. O'Connell, I. Osband, and I. C. Gilaberte. OpenAI o1 System Card. *CoRR*, abs/2412.16720, 2024.

D. Knuth. Claude's cycles. 2026.

E. Korkmaz. Deep reinforcement learning policies learn shared adversarial features across mdps. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 7229–7238. AAAI Press, 2022.

E. Korkmaz. Adversarial robust deep reinforcement learning requires redefining robustness. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 8369–8377. AAAI Press, 2023.

E. Korkmaz. Understanding and diagnosing deep reinforcement learning. In *Proceedings of the 41th International Conference on Machine Learning, ICML 2024*, 2024.

E. Korkmaz. Counteractive RL: Rethinking Core Principles for Efficient and Scalable Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025 [Spotlight Presentation]*, 2025.

E. Korkmaz. How to lose inherent counterfactuality in reinforcement learning. In *The Fourteenth International Conference on Learning Representations*. ICLR, 2026.

E. Korkmaz and J. Brown-Cohen. Detecting adversarial directions in deep reinforcement learning to make robust decisions. In *Proceedings of the 40th International Conference on Machine Learning, ICML 2023*, volume 202, pages 17534–17543, 2023.

A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, a. G. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

R. Montague, P. Dayan, and T. Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. In *Journal of Neuroscience*, 1996.

OpenAI. The Quasi-Riemann Hypothesis, 2026a.

OpenAI. Sharing AI progress in mathematics, 2026b.

L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta. Robust adversarial reinforcement learning. *International Conference on Learning Representations ICLR*, 2017.

T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016*, 2016.

J. Schmidhuber. Adaptive confidence and adaptive curiosity. In *Technical Report FKI*, 1991.

K. J. Seymour, T. Stein, C. W. Clifford, and P. Sterzer. Cortical suppression in human primary visual cortex predicts individual differences in illusory tilt perception. *Journal of Vision*, 2018.

S. Singh and P. Dayan. Analytical mean squared error curves for temporal difference learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 1996.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

G. Tesauro. Practical issues in temporal difference learning. In *Advances in Neural Information Processing Systems 4*, 1991.

The White House. Blueprint for an ai bill of rights. 2023.

United Kingdom Parliament. Artificial intelligence: ethics, governance and regulation. http://post.parliament.uk/artificial-intelligence-ethics-governance-and-regulation/, 2024.

H. van Hasselt. Double q-learning. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 2613–2621. Curran Associates, Inc., 2010.

H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2094–2100. AAAI Press, 2016.

Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas. Dueling network architectures for deep reinforcement learning. *Internation Conference on Machine Learning ICML.*, page 1995–2003, 2016.

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

G. Westheimer. Illusions in the spatial sense of the eye: Geometrical–optical illusions and the neural representation of space. *Vision Research*, 2008.

M. Yan. Extension of convex function. *Journal of Convex Analysis*, 21(4):965–987, 2014.

A Directions of Incoherence

We provided the Proposition 3.3 in Section 3 in the main body of the paper. Here we provide the full proof.

Proposition A.1 (*Directions of Incoherence*)**.** *Let $\pi_{\mathcal{Q}}(s,a) = \frac{\exp \mathcal{Q}(s,a)}{\sum_{a' \in \mathcal{A}} \exp \mathcal{Q}(s,a')}$ be the softmax policy defined by $\mathcal{Q}$. Let $a^*(\hat{s}) = \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a)$. Then the gradient of the cross-entropy loss $\mathcal{J}(s,\hat{s})$ is given by*

$$\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} = \mathbb{E}_{a \sim \pi_{\mathcal{Q}}(\hat{s},a)}[\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}] - \nabla_s \mathcal{Q}(s,a^*(\hat{s})|_{s=\hat{s}}]$$

Proof. Since the policy coherence loss is

$$\mathcal{J}(s,\hat{s}) = -\log \pi_{\mathcal{Q}}(s, \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a))$$

The gradient of the directions of incoherence with respect to states is

$$\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} = -\nabla_s \big[\log \pi_{\mathcal{Q}}(s, \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a))\big]|_{s=\hat{s}}$$

Since the softmax policy is $\pi_{\mathcal{Q}}(s,a) = \frac{\exp \mathcal{Q}(s,a)}{\sum_{a' \in \mathcal{A}} \exp \mathcal{Q}(s,a')}$.

$$\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} = -\nabla_s [\log \frac{\exp \mathcal{Q}(s, \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a))}{\sum_{a' \in \mathcal{A}} \exp \mathcal{Q}(s,a')}]|_{s=\hat{s}}$$
$$\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} = \nabla_s \big[\log \sum_{a' \in \mathcal{A}} e^{\mathcal{Q}(s,a')} - \mathcal{Q}(s, \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a))\big]$$

By the chain rule,

$$\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} = \frac{1}{\sum_{a' \in \mathcal{A}} e^{\mathcal{Q}(s,a')}} \sum_{a' \in \mathcal{A}} \nabla_s e^{\mathcal{Q}(s,a')} - \nabla_s \mathcal{Q}(s, \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a))$$

Hence,

$$\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} = \mathbb{E}_{a \sim \pi_{\mathcal{Q}}(\hat{s},a)}[\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}] - \nabla_s \mathcal{Q}(s, \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a)|_{s=\hat{s}}]$$

Note that $a^*(\hat{s}) = \arg\max_{a \in \mathcal{A}} \mathcal{Q}(\hat{s},a)$. Thus,

$$\nabla_s \mathcal{J}(s,\hat{s})|_{s=\hat{s}} = \mathbb{E}_{a \sim \pi_{\mathcal{Q}}(\hat{s},a)}[\nabla_s \mathcal{Q}(s,a)|_{s=\hat{s}}] - \nabla_s \mathcal{Q}(s,a^*(\hat{s})|_{s=\hat{s}}]$$

□

B Strict Convexity of Negative Average Advantage

Theorem B.1 (*Strict Convexity of Negative Average Advantage*)**.** *Let S^π be a finite sequence of states by the greedy policy π with respect to state-action value function $\mathcal{Q}$ in an MDP M. Assume that the state space $\mathcal{S}$ has dimension $n \geq |S^\pi|$, and that $\Omega(s,\hat{s})$ is a twice continuously differentiable function of s and strictly convex on the simplex $\mathrm{conv}(S^\pi)$. Then Ω can be extended to a strictly convex function on the entire state space.*

Proof. Observe that since S^π is a finite set of cardinality at most n, we have that $\mathrm{conv}(S^\pi)$ is compact. Since Ω is twice continuously differentiable, strict convexity of Ω on $\mathrm{conv}(S^\pi)$ implies that the Hessian is positive definite. Therefore, by Theorem 3.6 in the main body of the paper Ω can be extended to a twice continuously differentiable function with positive definite Hessian on the entire state space. Positive definiteness of the Hessian then implies that the extention of Ω to the entire state-space is strictly convex. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the discovery of intrinsic alignment between the advantage function and policy incoherence gradients. These claims are supported by theoretical proofs in Section 3 and empirical verification in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our work discovers an intrinsic alignment of directions of incoherence and the average advantage function. We then propose an algorithm to leverage this intrinsic alignment and audit the uncertainties in the environment. From algorithmic differences and architectural changes to distributional shift and worst-case changes in the environment our proposed algorithm, i.e. SCALE, can in fact diagnose and identify these changes.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each of the theoretical results provided with the full set of assumptions and a complete and correct proof. Theoretical analysis is reported in Section 3 with full proofs and assumptions reported in the same section 3. The proof of Proposition 3.3 is provided in the supplementary material. All the theorems, formulas, and proofs in our paper is numbered and cross-referenced.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental results are reported in Section 4 which further extends to Section 4.1, Section 4.3, and Section 4.4 with proper metrics explained in detail with algorithms that properly referenced and the hyperparameters for the methods used in 4.3, and Section 4.4 also explained in Section 4. The learning environment, i.e. data, is the Arcade Learning Environment, which is publicly available canonical training environment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The learning environment is the Arcade Learning environment which is already publicly available. The details regarding the experimental analysis is reported in Section 4 which further extends to Section 4.1, Section 4.3, and Section 4.4 with proper metrics explained in detail with algortihms that properly referenced and the hyperparameters for the methods used in 4.3, and Section 4.4 also explained in Section 4.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The details regarding the experimental analysis is reported in Section 4 which further extends to Section 4.1, Section 4.3, and Section 4.4 with proper metrics explained in detail with algortihms that properly referenced and the hyperparameters for the methods used in 4.3, and Section 4.4 also explained in Section 4. The hyperparameters for the trained policies are precisely the same as the papers proposed these algortihms to provide a transparent comparison and analysis. This is also stated in the main paper in section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results reported in our paper provides appropriate information about the statistical significance. The results are reported with standard error of the mean.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our work discovers a foundational relationship between average advantage and the directions of incoherence. The hyperparameters are shared in Section 4 to reproduce the results. Directions of incoherence are the canonical formulations that are referenced properly.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work does not violate the NeurIPS Code of Ethics in any form or aspect.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work provides foundational analysis on the functional structure of deep reinforcement learning manifold in high-dimensional MDPs. Furthermore, we propose an algorithm that can diagnose the environment the policy is interacting. Our results demonstrate that from algorithmic and architectural changes to adversarial perturbations and natural changes, our proposed algorithm SCALE can in fact audit differences. This diagnosis and identification of uncertainties is a concrete step towards achieving more stable, robust and generalizable deep reinforcement learning policies with responsible decision making capabilities. As the capabilities of artificial intelligence increases the risks associated with it rise as well where the regulatory policies and measures is currently being developed European Union Parliament [2023], The White House [2023], Europen Union [2024], United Kingdom Parliament [2024]. Our paper contributes to this axis with regard to providing principled analysis and actionable guidance on the diagnosis, transparent characterization and comprehensive evaluation of frontier AI systems.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work does not posit such a risk.

    Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The learning environment is the publicly available cononical learning environment that is referenced in section 4. The rest of the details are also properly referenced in Section 4.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The learning environment is the publicly available learning environment learning environment, i.e. Arcade Learning Environment.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Our work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Our paper does not involve any LLM use in any form.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.