Understanding and Diagnosing Deep Reinforcement Learning

Ezgi Korkmaz [1]

Abstract

Deep neural policies have recently been installed in a diverse range of settings, from biotechnology to automated financial systems. However, the utilization of deep neural networks to approximate the value function leads to concerns on the decision boundary stability, in particular, with regard to the sensitivity of policy decision making to indiscernible, non-robust features due to highly non-convex and complex deep neural manifolds. These concerns constitute an obstruction to understanding the reasoning made by deep neural policies, and their foundational limitations. Hence, it is crucial to develop techniques that aim to understand the sensitivities in the learnt representations of neural network policies. To achieve this we introduce a theoretically founded method that provides a systematic analysis of the unstable directions in the deep neural policy decision boundary across both time and space. Through experiments in the Arcade Learning Environment (ALE), we demonstrate the effectiveness of our technique for identifying correlated directions of instability, and for measuring how sample shifts remold the set of sensitive directions in the neural policy landscape. Most importantly, we demonstrate that state-of-the-art robust training techniques yield learning of disjoint unstable directions, with dramatically larger oscillations over time, when compared to standard training. We believe our results reveal the fundamental properties of the decision process made by reinforcement learning policies, and can help in constructing reliable and robust deep neural policies.

1. Introduction

Reinforcement learning algorithms leveraging the power of deep neural networks have obtained state-of-the-art results initially in game-playing tasks (Mnih et al., 2015) and subsequently in continuous control (Lillicrap et al., 2015). Since this initial success, there has been a continuous stream of developments both of new algorithms (Mnih et al., 2016; Hasselt et al., 2016; Wang et al., 2016), and striking new performance records in highly complex tasks (Silver et al., 2017; Schrittwieser et al., 2020). While the field of deep reinforcement learning has developed rapidly (Mankowitz et al., 2023), the understanding of the representations learned by deep neural network policies has lagged behind.

The lack of understanding of deep neural policies is of critical importance in the context of the sensitivities of policy decisions to imperceptible, non-robust features. Beginning with the work of (Szegedy et al., 2014; Goodfellow et al., 2015), deep neural networks have been shown to be vulnerable to adversarial perturbations below the level of human perception. In response, a line of work has focused on proposing training techniques to increase robustness by applying these perturbations to the input of deep neural networks during training time (i.e. adversarial training) (Goodfellow et al., 2015; Madry et al., 2017). Yet, concerns have been raised on these methods including decreased accuracy on clean data (Bhagoji et al., 2019), prohibiting generalization (Korkmaz, 2023), and incorrect invariance to semantically meaningful changes (Tramèr et al., 2020). While some studies argued that detecting adversarial directions could be the best we can do so far (Korkmaz & Brown-Cohen, 2023), the diagnostic perspective on understanding policy decision making and vulnerabilities requires urgent further attention.

Thus, it is crucial to develop techniques to precisely understand and diagnose the sensitivities of deep neural policies, in order to effectively evaluate newly proposed algorithms and training methods. In particular, there is a need to have diagnostic methods that can automatically identify policy sensitivities and instabilities that arise under many different scenarios, without requiring extensive research effort for each new instance.

For this reason, in our paper we focus on understanding the learned representations and policy vulnerabilities and ask the following questions: *(i) How can we analyze the rationale behind deep reinforcement learning decisions? (ii) What is the temporal and spatial relation between non-robust directions on the deep neural policy manifold? (iii)*

[1] University College London (UCL). Correspondence to: Ezgi Korkmaz <ezgikorkmazmail@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

How do the directions of instabilities in the deep neural policy landscape transform under a portfolio of state-of-the-art adversarial attacks? (iv) How does distributional shift affect the learnt non-robust representations in reinforcement learning with high dimensional state representation MDPs? (v) Does the state-of-the-art certified adversarial training solve the problem of learning correlated non-robust representations in sequential decision making? To be able to answer these questions in our paper from worst-case to natural directions we focus on understanding the representations learned by deep reinforcement learning policies and make the following contributions:

- We introduce a theoretically founded novel approach to systematically discover and analyze the spatial and temporal correlation of directions of instability on the deep reinforcement learning manifold.
- We highlight the connection between neural processing with visual illusion stimulus and our analysis to understand and diagnose deep neural policies. We conduct extensive experiments in the Arcade Learning Environment with neural policies trained in high-dimensional state representations, and provide an analysis on a portfolio of state-of-the-art adversarial attack techniques. Our results demonstrate the precise effects of adversarial attacks on the non-robust features learned by the policy.
- We investigate the effects of distributional shift on the correlated vulnerable representation patterns learned by deep reinforcement learning policies to provide a comprehensive and systematic robustness analysis of deep neural policies.
- Finally, our results demonstrate the presence of non-robust features in adversarially trained deep reinforcement learning policies, and that the state-of-the-art certified robust training methods lead to learning disjoint and spikier vulnerable representations.

2. Background and Preliminaries

2.1. Preliminaries

A Markov Decision Process (MDP) is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the Markov transition kernel, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. A reinforcement learning agent interacts with an MDP by observing the current state $s \in \mathcal{S}$ and taking an action $a \in \mathcal{A}$. The agent then transitions to state s' with probability $\mathcal{P}(s, a, s')$ and receives reward $\mathcal{R}(s, a, s')$. A policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ selects action a in state s with probability $\pi(s, a)$. The main objective in reinforcement learning is to learn a policy π which maximizes the expected cumulative discounted rewards $R = \mathbb{E}_{a_t \sim \pi(s_t, \cdot)} \sum_t \gamma^t \mathcal{R}(s_t, a_t, s_{t+1})$. This maximization is achieved by iterative Bellman update to learn a state-action value function (Watkins & Dayan, 1992)

$$Q(s_t, a_t) = \mathcal{R}(s_t, a_t, s_{t+1}) + \gamma \sum_{s_t} \mathcal{P}(s_{t+1}|s_t, a_t) V(s_{t+1}).$$

$Q(s, a)$ converges to the optimal state-action value function, representing the expected cumulative discounted rewards obtained by the optimal policy when starting in state s and taking action a, with value function $V(s) = \max_{a \in \mathcal{A}} Q(s, a)$. Hence, the optimal policy $\pi^*(s, a)$ can be obtained by executing the action $a^*(s) = \arg\max_a Q(s, a)$ i.e. the action maximizing the state-action value function in state s.

2.2. Adversarial Perturbation Techniques and Formulations

Following the initial study conducted by Szegedy et al. (2014), Goodfellow et al. (2015) proposed a fast and efficient way to produce ϵ-bounded adversarial perturbations in image classification based on linearization of $J(x, y)$, the cost function used to train the network, at data point x with label y. Consequently, Kurakin et al. (2016) proposed the iterative form of this algorithm: the iterative fast gradient sign method (I-FGSM).

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

This algorithm further has been improved by the proposal of the utilization of the momentum term (Dong et al., 2018). Following this Korkmaz (2020) proposed a Nesterov momentum technique to compute ϵ-bounded adversarial perturbations for deep reinforcement learning policies by computing the gradient at the point $s_{\text{adv}}^t + \mu \cdot v_t$,

$$v_{t+1} = \mu \cdot v_t + \frac{\nabla_{s_{\text{adv}}} J(s_{\text{adv}}^t + \mu \cdot v_t, a)}{\|\nabla_{s_{\text{adv}}} J(s_{\text{adv}}^t + \mu \cdot v_t, a)\|_1} \qquad (2)$$

$$s_{\text{adv}}^{t+1} = s_{\text{adv}}^t + \alpha \cdot \frac{v_{t+1}}{\|v_{t+1}\|_2} \qquad (3)$$

Another class of algorithms for computing adversarial perturbations focuses on different methods for computing the smallest possible perturbation which successfully changes the output of the target function. The DeepFool method of Moosavi-Dezfooli et al. (2016) works by repeatedly computing projections to the closest separating hyperplane of a linearization of the deep neural network at the current point. Carlini & Wagner (2017) proposed targeted adversarial formulations in image classification based on distance minimization between the original sample and the adversarial sample

$$\min_{x_{\text{adv}} \in \mathcal{X}} c \cdot J(x_{\text{adv}}) + \|x_{\text{adv}} - x\|_2^2 \qquad (4)$$

Another variant of this algorithm is based on ℓ_1-regularization of the ℓ_2-norm bounded Carlini & Wagner (2017) adversarial formulation (Chen et al., 2018).

$$\min_{x_{\text{adv}} \in \mathcal{X}} c \cdot J(x_{\text{adv}}) + \sigma_1 \left\| x_{\text{adv}} - x \right\|_1 + \sigma_2 \left\| x_{\text{adv}} - x \right\|_2^2$$

2.3. Deep Reinforcement Learning Policies and Adversarial Effects

Beginning with the work of Huang et al. (2017) and Kos & Song (2017), which introduced adversarial examples based on FGSM to deep reinforcement learning, there has been a long line of research on both adversarial attacks and robustness for deep neural policies. On the attack side, Korkmaz (2020) showed Nesterov momentum produced adversarial perturbations that are faster to compute compared to Carlini & Wagner (2017) with similar or better impact on the policy performance. More intriguingly, the work of Korkmaz (2022) discovered that deep reinforcement learning policies learn similar adversarial directions across MDPs intrinsic to the training environment, thus revealing an underlying approximately linear structure learnt by deep neural policies. On the defense side Pinto et al. (2017) model the interaction between an adversary producing perturbations and the deep neural policy taking actions as a zero-sum game, and train the policy jointly with the adversary in order to improve robustness. More recently, Huan et al. (2020) formalized the adversarial problem in deep reinforcement learning by introducing a modified MDP definition which they term State-Adversarial MDP (SA-MDP). Based on this model the authors proposed a theoretically motivated *certified robust* adversarial training algorithm called SA-DQN. Quite recently, Korkmaz (2023) provided a contrast between natural directions and adversarial directions with respect to their perceptual similarity to base states and impact on the policy performance. While the results in this paper demonstrate that certified adversarial training techniques limit the generalization capabilities of deep reinforcement learning policies, the paper further argues the need for rethinking robustness in deep reinforcement learning. While recent studies raised some concerns on the drawbacks of certified adversarial training techniques from generalization to security, these studies lack a method of explaining and understanding the main problems of robustness in deep reinforcement learning, and in particular with clear analysis of the vulnerabilities of the policies.

3. Probing the Deep Neural Policy Manifold via Non-Lipschitz Directions

In our paper our goal is to seek answers for the following questions:

- *What is the reasoning behind deep reinforcement learning decision making?*
- *How can we analyze the robustness of deep reinforcement learning policies across time and space?*
- *What are the effects of distributional shift on the vulnerable representations learnt?*
- *How do adversarial attacks remold the volatile patterns learnt by the neural policies?*
- *Does adversarial training ensure robust and safe policies without learning any non-robust features?*

To be able to answer these questions we propose a principled robustness appraisal method that probes the deep reinforcement learning manifold via non-Lipschitz directions across time and across space. In the remainder of this section we explain in detail our proposed method.

Definition 3.1 (ϵ*-non-Lipschitz Direction*)**.** Let Q be a state-action value function and let $\epsilon > 0$. For a state $s \in \mathcal{S}$ and vector $w \in \mathbb{R}^d$, let $\hat{s} = s + \epsilon w$. The vector v is an ϵ-non-Lipschitzness direction that uncovers the high-sensitivities of the deep neural manifold for Q in state s if

$$v = \underset{\|w\|_2 = 1}{\operatorname{argmax}}\, Q(\hat{s}, \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q(\hat{s}, a)) - Q(\hat{s}, \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q(s, a)). \tag{5}$$

In words, v is a non-Lipschitz direction when adding a perturbation of ℓ_2-norm ϵ along v maximizes the difference between the maximum state-action value in the new state and the value assigned in the new state to the previously maximal action. Eqn 5 can be approximated by using the softmax cross entropy loss.[1] The cross entropy loss between the softmax policy in state s_g and the argmax policy $\tau(s,a) = \mathbb{1}_{a = \operatorname{argmax}_{a'} \pi(s,a')}(a)$ at state s is

$$\begin{aligned} J(s, s_g) &= -\sum_{a \in \mathcal{A}} \tau(s,a) \log(\pi(s_g, a)) \\ &= -\log(\pi(s_g, a^*(s))). \end{aligned}$$

Therefore by definition of the softmax policy we have

$$\begin{aligned} J(s, s_g) &= \log \sum_{a' \in \mathcal{A}} e^{Q(s_g, a')/T} - Q(s_g, a^*(s))/T \\ &\approx (Q(s_g, a^*(s_g)) - Q(s_g, a^*(s)))/T \end{aligned}$$

where the final approximate equality becomes close to an equality as T gets smaller. Setting $v = s_g - s$, shows that maximizing the softmax cross entropy approximates the maximization in Definition 5. Hence, the gradient $\nabla_{s_g} J(s, s_g)|_{s_g = s}$ gives the direction of the largest increase in cross-entropy when moving from state s. Intuitively this

[1] $\pi(s,a)$ is defined as the softmax policy of the state-action value function $\pi(s,a) = \dfrac{e^{(Q(s,a)/T)}}{\sum_{a' \in \mathcal{A}} e^{(Q(s,a')/T)}}$.

is the direction along which the policy distribution $\pi(s, a)$ will most rapidly diverge from the argmax policy. Hence, $\nabla_{s_g} J(s, s_g)|_{s_g=s}$ is a high-sensitivity direction in the neural policy landscape in state s. Fundamentally, moving along the non-Lipschitz directions on the deep neural policy decision boundary will uncover the non-robust features learnt by the reinforcement learning policy. To capture the correlated non-robust features we must aggregate the information on high-sensitivity directions from a collection of states visited while utilizing the policy π in a given MDP. We thus define a single direction which captures the aggregate non-robust feature information from multiple states via the first principal component of the non-Lipschitz directions as follows:

Definition 3.2 (*Principal non-Lipschitz direction*)**.** Given a set of n states $S = \{s_i\}_{i=1}^n$ the principal non-Lipschitz direction is the vector $\mathcal{G}_S$ given by

$$\mathcal{G}_S = \underset{\{z \in \mathbb{R}^d \mid \|z\|_2 = 1\}}{\text{argmax}} \frac{1}{n} \sum_{i=1}^{n} \langle z, \nabla_{s_g} J(s_i, s_g)|_{s_g=s_i} \rangle^2.$$

Proposition 3.3 (*Spectral characterization of principal non-Lipschitz directions*)**.** *Given a set of n states $S = \{s_i\}_{i=1}^n$ define the matrix $\mathcal{L}(S)$ by*

$$\mathcal{L}(S) = \frac{1}{n} \sum_{i=1}^{n} \nabla_{s_g} J(s_i, s_g)|_{s_g=s_i} [\nabla_{s_g} J(s_i, s_g)|_{s_g=s_i}]^\top.$$

Then $\mathcal{G}_S$ is the eigenvector corresponding to the largest eigenvalue of $\mathcal{L}(S)$.

Proof. Observe that by linearity of the inner product

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \langle z, \nabla_{s_g} J(s_i, s_g)|_{s_g=s_i} \rangle^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} z^\top \nabla_{s_g} J(s_i, s_g)|_{s_g=s_i} [\nabla_{s_g} J(s_i, s_g)|_{s_g=s_i}]^\top z \\
&= z^\top (\frac{1}{n} \sum_{i=1}^{n} \nabla_{s_g} J(s_i, s_g)|_{s_g=s_i} [\nabla_{s_g} J(s_i, s_g)|_{s_g=s_i}]^\top) z \\
&= z^\top \mathcal{L}(S) z.
\end{aligned}$$

Thus $\mathcal{G}_S = \text{argmax}_{\{z \in \mathbb{R}^d \mid \|z\|_2 = 1\}} z^\top \mathcal{L}(S) z$. Therefore, by the variational characterization of eigenvalues, $\mathcal{G}_S$ is the eigenvector corresponding to the largest eigenvalue of $\mathcal{L}(S)$. □

Thus, the dominant eigenvector corresponds to $\mathcal{G}_S$, the largest correlation with non-Lipschitz directions across time, which follows from the standard analysis of principal component analysis. Also note that $\mathcal{G}_S$ has the same dimensions as each state s, and thus can easily be rendered in the same format as the states to visualize non-robust features. Proposition 3.3 shows that $\mathcal{G}_S$ can be computed by solving an eigenvalue problem. Proposition 3.3 is the basis for Algorithm 1, which computes $\mathcal{G}_S$ by first calculating $\mathcal{L}(S)$ by summing over states, and then outputs the maximum eigenvector. Next we demonstrate how RA-NLD can be used to measure the effects of environment changes on the correlated non-robust features both visually and quantitatively.

Algorithm 1 RA-NLD: Robustness Analysis via Non-Lipschitz Directions in the Deep Neural Policy Manifold

```
Input: MDP $\mathcal{M}$, state-action value function $Q(s, a)$, actions $a \in \mathcal{A}$, states $s \in \mathcal{S}$, the transition probability kernel $\mathcal{P}(s, a, s')$
Output: Principal non-Lipschitz direction $\mathcal{G}(i, j)$
for $s = s_0$ to $s_T$ do
    $\tau(s, a) = \mathbb{1}_{a=\text{argmax}_{a'} Q(s,a')}(a)$
    $\pi(s_\text{g}, a) = \text{softmax}(Q(s_\text{g}, a))$
    $J(s, s_g) = -\sum_{a \in \mathcal{A}} \tau(s, a) \log(\pi(s_g, a))$
    $\mathcal{L}$ += $\nabla_{s_g} J(s, s_g)|_{s_g=s} [\nabla_{s_g} J(s, s_g)|_{s_g=s}]^\top$
end for
Return: Eigenvector $\mathcal{G}$ corresponding to largest eigenvalue of $\mathcal{L}$
```

Definition 3.4 (*Encountered set of states*)**.** Let $\Psi : \mathcal{S} \to \mathcal{S}$ be a function that transforms states $s \in \mathcal{S}$ of an MDP $\mathcal{M}$. Let S be the set of states encountered when utilizing policy π in $\mathcal{M}$. Then S^Ψ is defined to be the set of states encountered when utilizing the policy $\pi \circ \Psi$ in $\mathcal{M}$ i.e. when the policy state observations are transformed via Ψ.

In this setting, comparing $\mathcal{G}_S$ and $\mathcal{G}_{S^\Psi}$ will provide a qualitative picture of how the environmental change affects the learned vulnerable representation patterns. In order to give a more quantitative metric for this change we define

Definition 3.5 (*Feature Correlation Quotient*)**.** For two sets of states S and S', the feature correlation quotient is given by

$$\Lambda(S', S) = \frac{\mathcal{G}_{S'}^\top \mathcal{L}(S) \mathcal{G}_{S'}}{\mathcal{G}_S^\top \mathcal{L}(S) \mathcal{G}_S}.$$

Proposition 3.6 (*Boundedness of Feature Correlation Quotient*)**.** *For any two sets of states S and S' it holds that* $0 \leq \Lambda(S', S) \leq 1$.

Proof. By Proposition 3.3,

$$\mathcal{G}_{S'}^\top \mathcal{L}(S) \mathcal{G}_{S'} \leq \max_{\|z\|_2=1} z^\top \mathcal{L}(S) z = \mathcal{G}_S^\top \mathcal{L}(S) \mathcal{G}_S$$

Thus the numerator of $\Lambda(S', S)$ is always less than or equal to the denominator i.e. $\Lambda(S', S) \leq 1$. Furthermore, $\mathcal{L}(S)$ is positive semidefinite, as it is a sum of rank one projection matrices, and hence $\Lambda(S', S) \geq 0$. □

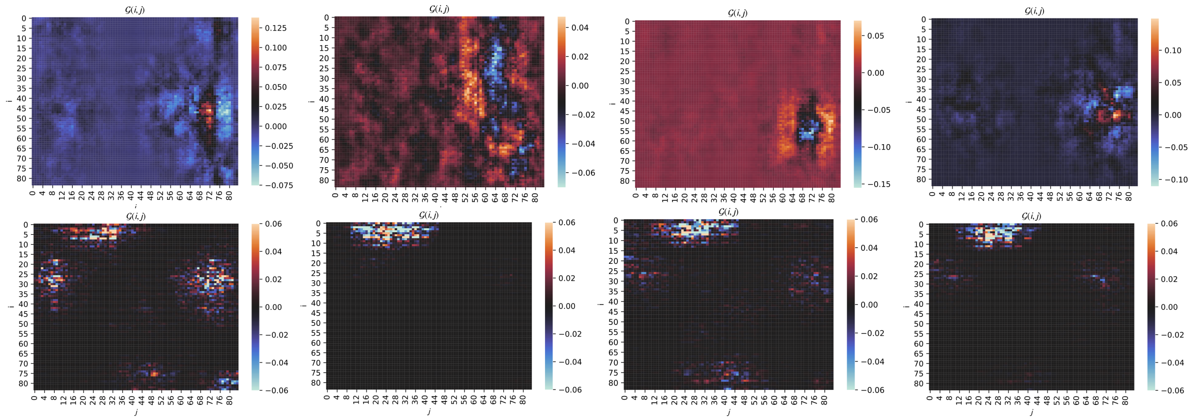


Figure 1. RA-NLD results of untransformed states and states under adversarial perturbations computed via Carlini&Wagner, Nesterov Momentum, and elastic-net regularization for Pong and BankHeist. Row1: Pong. Row2: BankHeist. Column1: Untransformed. Column2: C&W. Column3: Nesterov Momentum. Column4: Elastic-Net

Table 1. The feature correlation quotient $\Lambda(\hat{S}, S)$ and $\Lambda(S^{\Psi}, S)$ for the adversarial transformations: Carlini&Wagner, Nesterov Momentum, DeepFool, Elastic-Net.

Environments	Base Observations	Carlini&Wagner	Nesterov Momentum	DeepFool	Elastic-Net
Freeway	0.9917±0.0023	0.9499±0.02056	0.7868±0.02162	0.6869±0.02981	0.72590±0.0592
BankHeist	0.8360±0.0116	0.2837±0.02316	0.3407±0.02412	0.1748±0.04421	0.30917± 0.0521
RoadRunner	0.7652±0.0385	0.1621±0.02199	0.3826±0.03118	0.5353±0.03127	0.52506± 0.0782
Pong	0.4934±0.0391	0.0408±0.04056	0.3444±0.01981	0.3277±0.02871	0.10529± 0.0629

Therefore, the feature correlation quotient $\Lambda(S', S)$ is a number between zero and one which intuitively measures how correlated the non-robust features from S' are to those from S. When measuring how an environmental change affects the decisions made by the deep neural policy and the non-robust representations learnt, it is also important to take the stochastic nature of the MDP into account. In particular, the non-robust features observed with two different executions of the same policy may differ slightly due to the inherent randomness of the MDP. To account for this, we first collect a baseline set of states with no modification S. We then collect a set of states $\hat{S}$ with no modification, and S^{Ψ} with modification. By comparing $\Lambda(\hat{S}, S)$ to $\Lambda(S^{\Psi}, S)$ we can see how much of the decrease in average correlation is caused by the stochastic nature of the MDP, and how much of the decrease is caused by the environmental change.

4. Experimental Analysis

The deep reinforcement learning policies evaluated in our experiments are trained with the Double Deep Q-Network algorithm (Hasselt et al., 2016) initially proposed in (van Hasselt, 2010) with the architecture proposed by Wang et al. (2016), and State-Adversarial Double Deep Q-Network (see Section 2.3) with experience replay (Schaul et al., 2016). The set of states S is collected over 10 episodes. We use the adversarial methodology from Korkmaz & Brown-Cohen (2023). The adversarial perturbation hyperparameters are: for the Carlini&Wagner formulation κ is 10, learning rate is 0.01, initial constant is 10, for the elastic-net regularization formulation β is 0.0001, learning rate is 0.1, maximum iteration is 300, for Nesterov Momentum ϵ is 0.001, and decay factor is 0.1.[2]

4.1. Non-Robust Feature Shifts under Adversarial Perturbations

In this section we investigate the effects of adversarial attacks on the learnt correlated non-robust features. Figure 1 reports the RA-NLD results for the untransformed states and the adversarially attacked state observations. In particular, these perturbations are computed via the Nesterov momentum, Carlini&Wagner, and elastic-net regularization formulations (see Section 2.2). Figure 1 demonstrates that different adversarial formulations surface different sets of correlated non-robust features. Depending on the perturbation type, the correlated directions of instability can change quite noticeably. In fact, while the Carlini&Wagner formulation leaves a distinct signature on the vulnerable representation pattern, the non-robust features under Nesterov

[2] The hyperparameters for the adversarial attacks are fixed to the same levels as base studies to provide transparency and consistency with the prior work. Furthermore, note that the setting is also optimized to achieve the most effective adversarial perturbations (i.e. perturbations causing the largest decrease on the discounted expected cumulative rewards obtained by the policy).

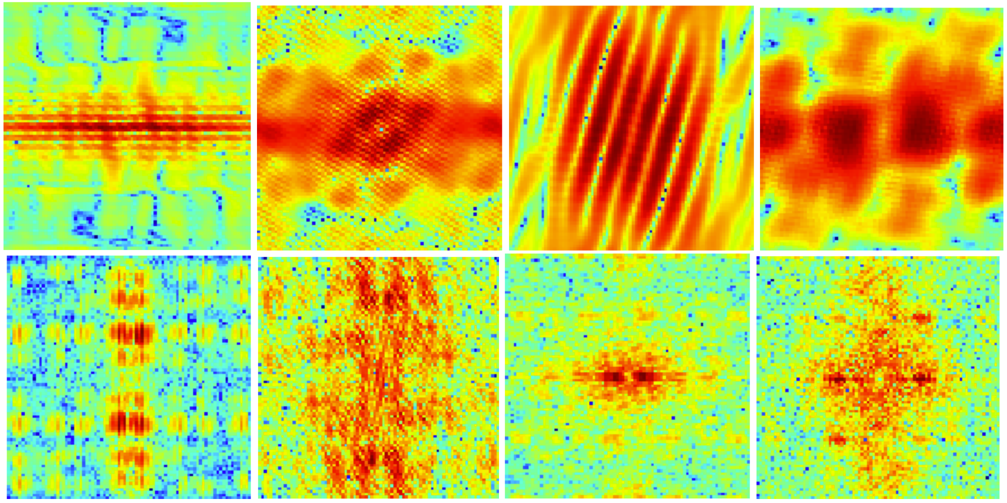


Figure 2. Fourier spectrum of the RA-NLD of the state-of-the-art adversarially and vanilla trained deep neural policies.[3] Row1: Adversarial. Row2: Vanilla. Column1: RoadRunner. Column2: BankHeist. Column3: Pong. Column4: Freeway

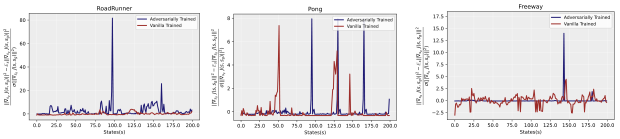
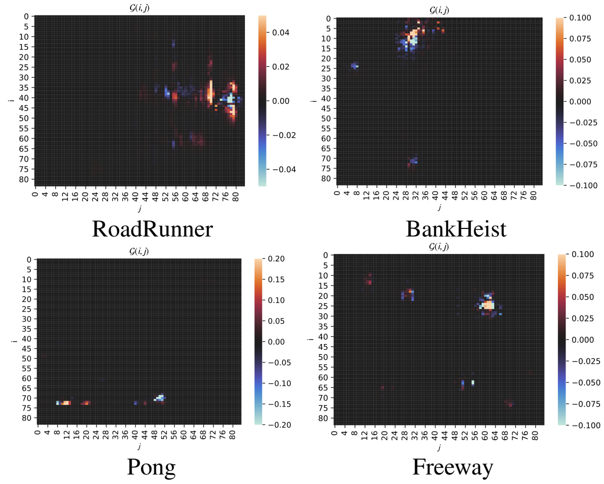


Figure 3. Standardized gradients $\|\nabla_{s_g} J(s_i, s_g)\|^2$ for vanilla trained and state-of-the-art certified adversarially trained deep reinforcement learning policies.

momentum appear most similar to those of the untransformed states. Thus, evidently our imaging technique helps to understand the rationale behind policy decision making and the vulnerabilities of deep reinforcement learning policies by allowing us to visualize precisely how non-robust features change with different sets of specifically optimized adversarial directions. Table 1 reports the feature correlation quotient $\Lambda(\hat{S}, S)$ and $\Lambda(S^{\Psi}, S)$ results where S consists of untransformed states and S^{Ψ} consists of states modified by the Nesterov Momentum, Carlini&Wagner, elastic-net regularization and DeepFool formulations respectively. Note that in all games the setting where $\hat{S}$ consists of a set of untransformed states from an independent execution has the highest feature correlation quotient $\Lambda(\hat{S}, S)$. Therefore the additional decrease of $\Lambda(S^{\Psi}, S)$ when S^{Ψ} is modified by adversarial perturbations can be attributed to changes in non-robust features caused by the perturbations. Observe also that the qualitative similarity between the visualizations in Figure 1 of the different transformed states is matched by their ranking under $\Lambda(S^{\Psi}, S)$ i.e. sorting from largest to smallest correlation quotient for BankHeist yields Nesterov momentum, Elastic-Net, and then Carlini&Wagner. The fact that the feature correlation quotient has distinct results for untransformed states and for states under all the types of adversarial formulations indicates that RA-NLD can facilitate detecting different types of adversarial perturbations. Measuring stimulus response to visual illusions has been used as an analysis tool in neural processing (Hubel & Wiesel, 1962; Grunewald & Lankheet, 1996; Westheimer, 2008; Seymour et al., 2018). One way to understand our approach is to examine the studies that focus on investigating the cortical area, parahippocampal cortex and hippocampus against visual illusion stimulus (Grunewald & Lankheet, 1996; Axelrod et al., 2017).



RoadRunner

BankHeist

Pong

Freeway

Figure 4. Principal non-Lipschitz direction $\mathcal{G}(i, j)$ for the state-of-the-art certified adversarially trained deep reinforcement learning policies for BankHeist, Pong, Freeway and RoadRunner.

[3] Figure 2 reports the Fourier transform of $\mathcal{G}_S$ where S is collected from a vanilla and adversarially trained policies in RoadRunner, BankHeist, Pong and Freeway. The Fourier transform reveals clear differences in the spatial frequencies occupied by $\mathcal{G}_S$ under vanilla and adversarial training. There is a consistent trend that the larger entries of the Fourier transform are more evenly and smoothly spread out for the adversarially trained policies. Thus, adversarial training leaves a consistent signature on the non-robust features detectable via the Fourier transform of $\mathcal{G}_S$. There is also a change in orientation: if the larger entries of the Fourier transform for the vanilla trained policy are more spread out along one axis,

Table 2. The feature correlation quotient $\Lambda(S', S)$ in BankHeist, Freeway, RoadRunner, and Pong for the natural transformations: brightness and contrast, compression artifacts, rotation modification, perspective transform, blurred observations.

Distributional Shift	Freeway	BankHeist	RoadRunner	Pong
Untransformed States	0.9917±0.0023	0.8360±0.0116	0.7652±0.0385	0.4934±0.0391
Brightness and Contrast	0.86756±0.0271	0.3095±0.0429	0.4369±0.0334	0.1678±0.0427
Compression Artifacts	0.90564±0.237	0.38814±0.022	0.24358±0.0204	0.49341±0.0191
Rotation Modification	0.1381±0.0081	0.2951±0.0062	0.3350±0.0050	0.13648±0.0032
Perspective Transform	0.3010±0.0281	0.1723±0.0311	0.3308±0.0274	0.4278±0.0196
Blurred Observations	0.2657±0.0148	0.0954±0.0127	0.2496±0.0162	0.0847±0.0083

4.2. Vulnerable Representations Learnt via Certified Adversarial Training

In this section we investigate the effects of adversarial training on the correlated non-robust features. In particular, the SA-DDQN algorithm adds the regularizer $\mathcal{R}$,

$$\mathcal{R}(\theta) = \sum_{s} \left(\max_{\bar{s} \in D_{\epsilon}(s)} \max_{a \neq a^*(s)} Q_{\theta}(\bar{s}, a) - Q_{\theta}(\bar{s}, a^*(s)) \right).$$

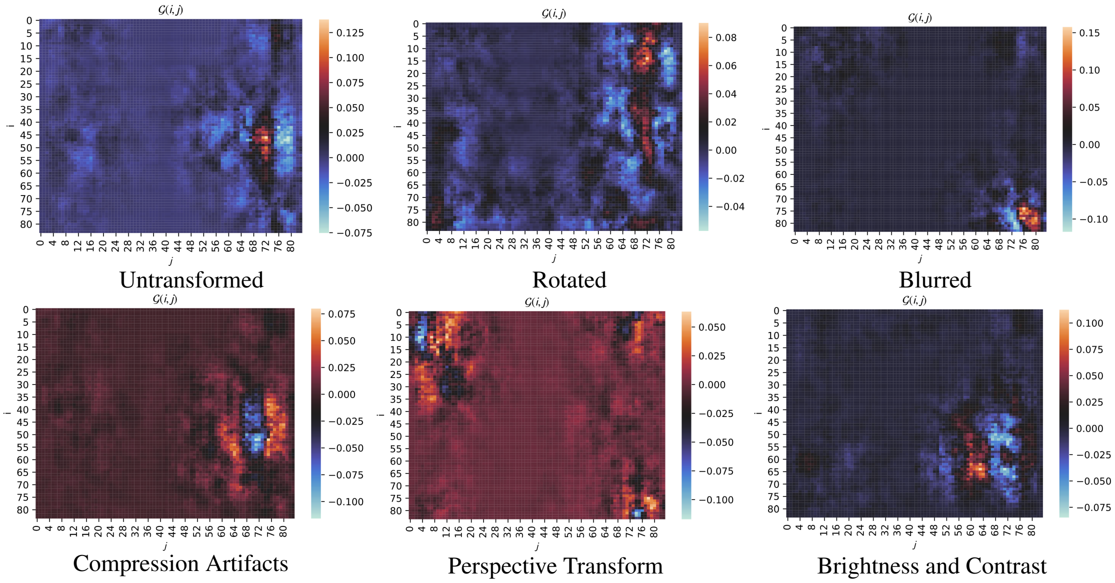
during training in the temporal difference loss. Figure 4 shows the RA-NLD results for the state-of-the-art adversarially trained deep reinforcement learning policies. The non-robust features of the adversarially trained deep neural policies are much more tightly concentrated on disjoint coordinates in the state observations, and these areas of concentration have moved significantly from where they were under vanilla training. Thus, the visualization allows us to see that correlated, non-robust features persist in adversarially trained policies, albeit in different locations with disjoint patterns than vanilla trained deep reinforcement learning policies. To complete our analysis of adversarial training we further include results on how non-robust features vary across time. For this purpose the ℓ_2-norm of the gradient $\|\nabla_{s_g} J(s_i, s_g)\|^2$ in each state $s_i \in S$ is recorded for both adversarially trained and vanilla trained policies in RoadRunner, Pong, and Freeway. The results are plotted in Figure 3. In both RoadRunner and Freeway, the adversarially trained policy has much higher variance in the gradient norm and thus in the level of instability. This is in contrast to the vanilla trained policy which tends to have a much smoother distribution which remains closer to the mean. These results indicate that adversarial training introduces higher jumps in sensitivity over states (i.e. extreme instability) when compared to vanilla training.

4.3. The Effects of Imperceptible Distributional Shift on the Directions of Instabilities

To evaluate the effects of distributional shift on the learnt policy we provide analysis on several environment modifications with RA-NLD. These transformations are natural semantically meaningful changes to the given MDP that correspond to imperceptible modifications to the state observations. In particular, the imperceptibility $\mathcal{P}_{\text{similarity}}$ is measured by, $\mathcal{P}_{\text{similarity}}(s, \Psi(s)) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}^l_{shw} - \hat{y}^l_{\Psi(s)hw})\|^2_2$ where $\hat{y}^l_s, \hat{y}^l_{\Psi(s)} \in \mathbb{R}^{W_l \times H_l \times C_l}$ represent the vector of unit normalized activations in the convolutional layers with width W_l, height H_l, and C_l is the number of channels.[4] Figure 5 reports $\mathcal{G}_S$ for states S collected under the six environment modifications mentioned above. For the untransformed setting the visualization of $\mathcal{G}_S$ clearly emphasizes the center of the region where the agent's paddle moves up and down to hit the ball. The components of $\mathcal{G}_S$ take larger positive values at the center of this region and transition to negative values along the boundary. A similar emphasis can be found for the case of compression artifacts, but with the signs reversed (i.e. the center of the region is negative and the boundary is positive). The other transformations exhibit larger changes in the regions emphasized in the visualization with perspective transform, blurring, rotation, and B&C causing the emphasized region to move to different locations. Table 2 contains the values of $\Lambda(\hat{S}, S)$ and $\Lambda(S^{\Psi}, S)$ where S is collected from an untransformed run and S^{Ψ} is collected from each of the six different transformations. In every game the largest value of $\Lambda(\hat{S}, S)$ occurs when $\hat{S}$ comes from an independent untransformed run, indicating that the additional decrease observed for S^{Ψ} from transformed runs is caused by the respective environmental transformations. It is notable that in Pong the second highest value for $\Lambda(S^{\Psi}, S)$ occurs for S^{Ψ} collected with compression artifacts, as this corresponds precisely to the qualitative similarity between the regions emphasized in the visualization of $\mathcal{G}_S$ for untransformed and compression artifacts. Hence, the results for $\Lambda(S^{\Psi}, S)$ help us to quanti-

the adversarially trained Fourier transform is more spread along the other.

[4] These imperceptible transformations include perspective transform, blurring, rotation, brightness, contrast, and compression artifacts as proposed in Korkmaz (2023). In particular, brightness and contrast is given by linear transformation, and compression artifacts are the diminution in high frequency components due to JPEG conversion. Note that this recent work demonstrates that these natural imperceptible transformations cause more damage to the policy performance compared to adversarial perturbations, and further highlights that the certified adversarial training is more vulnerable towards these natural attacks.



Untransformed

Rotated

Blurred

Compression Artifacts

Perspective Transform

Brightness and Contrast

Figure 5. RA-NLD results of untransformed state observations and states under natural transformations with rotation, perspective transformation, blurring, compression artifacts, and B&C for Pong.

tatively understand the effects of the environmental changes in the MDP, while agreeing well with the qualitative results of the RA-NLD outputs.

4.4. RA-NLD to Understand Policy Decision Making and Diagnose Non-Robustness

By leveraging the non-Lipschitz direction analysis not only can we uncover non-robust representations learnt deep neural policies, further we can analyze how their decisions are formed given an MDP and a training algorithm and what makes these decisions change under different influences from adversarial manipulations and natural changes in a given environment. While the RA-NLD visualizations give us semantically meaningful information on how policy decisions are influenced and the non-robust features learnt by the deep neural policy, they also provide a detailed understanding of how these voaltile representations change under non-stationary MDPs. The fact that RA-NLD can provide fine-grained vulnerability analysis of deep reinforcement learning policies under adversarial attacks, with distributional shift and with different training algorithms can help with diagnosis of policy vulnerabilities in the development phase.

Conducting ablation studies with RA-NLD in reinforcement learning algorithm design can prevent building policies with inherent non-robustness, and our algorithm can be utilized to visualize and identify the effects of several design choices (e.g. algorithm, neural network architecture) on the non-robust features learnt by the policy from the MDP. In particular, given a visualization of the vulnerability pattern for a trained policy, one can try to modify the training environment in a way that will make the policy invariant to the non-robust features revealed by RA-NLD. Such modification could include changing the state representation in a way that does not change the semantics of the MDP or the task at hand, but does change the inherent non-robustness in question. Furthermore, the effect of modifications to training algorithms can also be directly visualized, as exemplified by our results for adversarial training. Thus our method gives a straightforward way to diagnose or debug any proposed methods in terms of their effects on the non-robustness of the neural policy and the volatile representations learnt by it.

One intriguing fact is that RA-NLD can uncover the vulnerable representations learnt by the certified adversarial training techniques. From the safety point of view it warrants significant concern that the algorithms targeting and certifying robustness end up learning non-robust representations. From the alignment perspective RA-NLD discovers that certified adversarial training is still producing misaligned deep reinforcement learning policies. Ultimately, for future research directions it is important to lay out exact trade-offs and vulnerabilities for these algorithms to eliminate the bias they can create for future research efforts. The impact of the imperceptible environmental changes in the MDP is immediately captured by the principal high-sensitivity direction analysis. The most intriguing aspect of these results is that not only can RA-NLD be used as a diagnostic tool during training, but further the principal non-Lipschitz direction analysis can also guide agents in real life on real-time understanding of the current rationale behind their decisions and their vulnerabilities. The RA-NLD algorithm gives us semantically meaningful information on the non-robust features learnt by the deep neural policy, and also provides a detailed understanding of how these non-robust features change under non-stationary environments.

5. Conclusion

In our paper we aim to seek answers for the following questions: *(i) How can we analyze the robustness and reliability of deep reinforcement learning policy decisions? (ii) What is the relation of non-robust representations learnt by deep neural policy temporally and spatially ? (iii) What are the effects of adversarial attacks on correlated non-robust features? (iv) Does adversarial training ensure safety and provide robust policies that do not learn non-robust representations? (v) How does distributional shift affect the learnt correlated non-robust features?* To be able to answer these questions we analyze non-Lipschitz directions in the deep neural policy landscape and we propose a novel technique to analyze and lay out correlated non-robust features learned by deep reinforcement learning policies. We show that deep reinforcement learning policies do end up learning correlated non-robust vulnerable representations, and that adversarial attacks lead to surfacing a new set of non-robust features or highlighting the existing ones. Most importantly, our results show that the state-of-the-art adversarial training techniques also end up learning temporally and spatially correlated non-robust features. Finally, we demonstrate that distributional shifts introduce different sets of correlated non-robust features compared to adversarial attacks. Hence, our analysis not only allows us to effectively visualize correlated directions of instability, but also allows for precise understanding of changes in the learnt non-robust representations caused by different training algorithms and different methods for altering states. Thus, we believe that our analysis can be critical both in understanding deep reinforcement learning policy decision making and in diagnosing the vulnerabilities of deep neural policies, while further enhancing our ability to design algorithms to improve robustness.

Impact Statement

The risks of artificial intelligence regarding safety have never been as prominent as they are in the current time (Tobin, 2023). From highly capable large language models (Google Gemini, 2023; OpenAI, 2023) to autonomous driving vehicles, these risks arise in real life (The New York Times, Decemeber 2023) as regulatory acts are being formed (The White House, 2023; European Comission, 2023; European Parliament, 2023). Our paper provides the necessary diagnostic tools to understand and interpret AI systems (i.e. deep reinforcement learning policies). Our paper introduces a theoretically founded technique to understand the vulnerabilities and volatilities of deep neural policies. Our results discover that *certified robust* training techniques have spikier volatilities resulting in revealing the current problems of safety guarantees in adversarial training techniques. We believe that it is crucial to understand the exact problems that might arise from the deep reinforcement learning policies before these policies are deployed in real life (The New York Times, 2022).

References

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

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 7496–7508, 2019.

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

Chen, P., Sharma, Y., Zhang, H., Yi, J., and Hsieh, C. EAD: elastic-net attacks to deep neural networks via adversarial examples. In McIlraith, S. A. and Weinberger, K. Q. (eds.), *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*, pp. 10–17. AAAI Press, 2018.

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

European Comission. Regulatory framework proposal on artificial intelligence. 2023.

European Parliament. EU AI act: First regulation on artificial intelligence. 2023.

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

Google Gemini. Gemini: A family of highly capable multimodal models. *Technical Report*, https://arxiv.org/abs/2312.11805, 2023.

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

Hasselt, H. v., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.

Huan, Z., Hongge, C., Chaowei, X., Li, B., Boning, M., Liu, D., and Hsiesh, C. Robust deep reinforcement learning against adversarial perturbations on state observatons. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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

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

Korkmaz, E. Nesterov momentum adversarial perturbations in the deep reinforcement learning domain. *International Conference on Machine Learning, ICML 2020, Inductive Biases, Invariances and Generalization in Reinforcement Learning Workshop.*, 2020.

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

Korkmaz, E. Adversarial robust deep reinforcement learning requires redefining robustness. *AAAI Conference on Artificial Intelligence*, 2023.

Korkmaz, E. and Brown-Cohen, J. Detecting adversarial directions in deep reinforcement learning. *International Conference on Machine Learning (ICML)*, 2023.

Kos, J. and Song, D. Delving into adversarial attacks on deep policies. *International Conference on Learning Representations*, 2017.

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

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXivpreprint arXiv:1509.02971*, 2015.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mankowitz, D. J., Michi, A., Zhernov, A., Gelmi, M., Selvi, M., Paduraru, C., Leurent, E., Iqbal, S., Lespiau, J., Ahern, A., Köppe, T., Millikin, K., Gaffney, S., Elster, S., Broshear, J., Gamble, C., Milan, K., Tung, R., Hwang, M., Cemgil, T., Barekatain, M., Li, Y., Mandhane, A., Hubert, T., Schrittwieser, J., Hassabis, D., Kohli, P., Riedmiller, M. A., Vinyals, O., and Silver, D. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964):257–263, 2023.

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

Mnih, V., Puigdomenech, A. B., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. *In International Conference on Machine Learning*, pp. 1928–1937, 2016.

Moosavi-Dezfooli, S., Fawzi, A., and Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2574–2582. IEEE Computer Society, 2016.

OpenAI. Gpt-4 technical report. *CoRR*, 2023.

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

Schaul, T., Quan, J., Antonogloua, I., and Silver, D. Prioritized experience replay. *International Conference on Learning Representations (ICLR)*, 2016.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T. P., and Silver, D. Mastering atari, go, chess and shogi by planning with a learned model. *Nat.*, 588(7839):604–609, 2020.

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

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, a., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Driessche, v. d. G., Graepel, T., and Hassabis, D. Mastering the game of go without human knowledge. *Nature*, 500:354–359, 2017.

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

The New York Times. Tesla autopilot and other driver-assist systems linked to hundreds of crashes. 2022.

The New York Times. The times sues openai and microsoft over a.i. use of copyrighted work. Decemeber 2023.

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

Tobin, J. Artificial intelligence: Development, risks and regulation. *United Kingdom Parliament*, 2023.

Tramèr, F., Behrmann, J., Carlini, N., Papernot, N., and Jacobsen, J. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9561–9571. PMLR, 2020.

van Hasselt, H. Double q-learning. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *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*, pp. 2613–2621. Curran Associates, Inc., 2010.

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

Watkins, C. J. C. H. and Dayan, P. Q-learning. 1992.

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