Illusionary Attacks on Sequential Decision Makers and Countermeasures

Tim Franzmeyer [1] **João F. Henriques** [1] **Jakob N. Foerster** [1] **Philip H.S. Torr** [1]
Adel Bibi [*1] **Christian Schroeder de Witt** [*1]

Abstract

Autonomous intelligent agents deployed to the real-world need to be robust against adversarial attacks on sensory inputs. Existing work in reinforcement learning focuses on *minimum-norm* perturbation attacks, which were originally introduced to mimic a notion of perceptual invariance in computer vision. In this paper, we note that such minimum-norm perturbation attacks can be trivially detected by victim agents, as these result in observation sequences that are not consistent with the victim agent's actions. Furthermore, many real-world agents, such as physical robots, commonly operate under human supervisors, which are not susceptible to such perturbation attacks. As a result, we propose to instead focus on *illusionary* attacks, a novel form of attack that is consistent with the *world model* of the victim agent. We provide a formal definition of this novel attack framework, explore its characteristics under a variety of conditions, and conclude that agents must seek *realism feedback* to be robust to illusionary attacks.

1. Introduction

Deep learning-based algorithms (Mnih et al., 2015; Schulman et al., 2017; Haarnoja et al., 2018; Salimans et al., 2017, DQN, PPO, SAC, ES) have found applications across a number of sequential decision making problems, ranging from simulated and real-world robotics (Todorov et al., 2012; Andrychowicz et al., 2020) to arcade games (Mnih et al., 2015) . It has recently been found, however, that deep neural network control policies conditioning on high-dimensional sensory input are prone to observation-space adversarial attacks, which poses threats to security and safety-critical applications (Kos & Song, 2017; Huang et al., 2017) and thus stimulates research into robust learning algorithms (zha, 2020).

Adversarial attacks on high-dimensional perceptual sensory inputs are commonly based on approaches that exploit universal shortcomings of neural network function approximators (Chaubey et al., 2020). In applications outside of sequential decision-making, the dominant school of thought is that adversarial perturbation vectors should be of *minimum norm* - either because it is thought that attacks carry a *cost* proportional to the perturbation size (Biggio et al., 2013), or by thinking of this requirement as a proxy to *human perceptual invariance* in image classification (Szegedy et al., 2013). Recent work applies the minimum-norm perturbation framework to deep reinforcement learning agents, both to identify possible attacks and defenses thereagainst (Lin et al., 2017; Chen et al., 2019; Sun et al., 2020)

We argue that the minimum-norm perturbation framework on its own is not adequate for adversarial attacks on sequential-decision makers, as minimum-norm attacks result in inconsistent action-observation sequences, and can be trivially detected by agents that possess a model of the environment. Powerful adversaries that seek to remain undetected will hence conduct *illusionary attacks*, which would perfectly replace a victim agent's reality by an internally coherent alternative one. Further, the framework of illusionary attacks generalizes beyond neural-network policies of autonomous agents and is feasible for adversarial attacks on human-in-the-loop settings, required for almost every enacted or emerging AI regulatory framework.

Illusionary attacks assume a threat model in which the adversary can inject perturbations into the victim agent's observations in order to trick it into executing a policy that is aligned with the adversary's objective. In practice, such perturbations could be achieved both through a software-level attack, or through direct operations on sensory input . The illuysionary attack framework requires perturbations to be *world-model aligned*, meaning that the resultant victim agent's action-observation histories need to be aligned with the model of the environment possessed by the agent, e.g., a model learned during training.

In practice, parts of an agent's observations may not be

[*]Equal supervision. [1]University of Oxford, UK. Correspondence to: Tim Franzmeyer <frtim at robots dot ox dot ac dot uk>.

Workshop on New Frontiers in Adversarial Machine Learning, International Conference on Machine Learning 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

accessible to the adversarial attacker. For example, a rescue robot's optical sensors may be perturbed by an adversary, but internal accelerometers and logging frameworks might be inaccessible. Such unperturbed parts of the victim agent's observations may thus form *realism feedback*. The existence of realism feedback channels therefore opens avenues for the victim agent to robustify its policy through preventative information gathering (Zintgraf et al., 2020).

In this paper, we proceed by first discussing related work (Section 2) and giving the necessary background (Section 3). We define our novel illusionary attack framework and the specialisations empirically investigated in this paper (Section 4.2). We analyse various illusionary attack settings empirically (Section 5), followed by a discussion and conclusion in Section 6.

2. Related Work

Adversarial attacks literature originates in non-sequential decision making applications such as image classification (Szegedy et al., 2013; Goodfellow et al., 2014), where the goal is to find perturbations δ for a given classifier f such that f yields different predictions for x and $x + \delta$, despite the difference between x and $x + \delta$ being imperceptible to humans. To enforce the imperceptibility requirement, such works enforce simple minimum-norm perturbations constraints (Goodfellow et al., 2014). **Adversarial robustness** aims at training networks that are robust against adversarial attacks. Among the most popular approaches are adversarial training (Madry et al., 2017) and randomized smoothing (Cohen et al., 2019).

Prior work on adversarial attacks on **sequential decision making agents** (Lin et al., 2017; Chen et al., 2019; Qiaoben et al., 2021) largely builds upon the minimum-norm perturbation framework. In particular, the observations of the agents are perturbed with minimum-norm attacks at every step. On another note, Hussenot et al. (2019) introduced a class of adversaries for which a unique mask is precomputed and added to the agents observation at every time step. Gleave et al. (2021) studied adversarial attacks in embodied multi-agent environments, defining the latter as a competitive multi-agent learning problem in which the adversary is part of the environment state, i.e. the adversary does not perturb the observation but affects the environment state.

Similarly to adversarial attacks on classification tasks, the minimum-norm perturbations framework has also been utilized towards building **robust agents in sequential decision environments**. For example, empirical robustness approaches were also proposed for this setting where approximations of worst case adversaries are augmented during policy training (Pattanaik et al., 2018). Moreover, certified methods for lower bound performance have also been proposed for sequential decision settings. For instance, Everett et al. (2021) proposed interval bound propagation through the network approximating the Q-function towards computing a certifiable lower bound on the cumulative reward. Moreover, following the recent advances in randomized smoothing, policy smoothing techniques were also shown to be effective in attaining provable lower bounds on the cumulative rewards, *i.e.* certifiably robust performance, (Kumar et al., 2021; Wu et al., 2021).

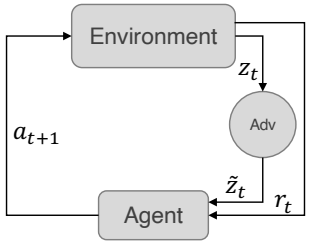

Figure 1: Test-time loop of adversarial attacks on agent observations.

We argue that definitions of minimum-norm perturbations are not adequate for sequential decision making agents as they can be easily detected and accordingly propose a new class of adversarial attacks, namely illusionary attacks. For completion, we will cover various robustness and detection approaches against minimum-norm perturbations in sequential decision settings.

Another extensive body of work has been focusing on **detecting adversarial attacks**. **?** develop an action-conditioned frame module that allows agents to detect adversarial attacks by comparing both the module's output distribution with the realised output distribution. Behzadan & Munir (2017) show that DQN (Mnih et al., 2015) agents under noncontiguous random perturbation training-time attacks can recover and adapt to the adversarial conditions by reactively by adjusting their policy. Havens et al. (2018) demonstrate how agents can detect adversairal attacks by comparing the attained reward to the estimated Q values, thereby assuming that the agent observes the unperturbed rewards. Tekgul et al. (2021) detect adversaries by evaluating the feasibility of past action sequences.

3. Background

3.1. Notations and setup – victim agent

POMDPs. We generally consider two agents: The victim agent and the adversary agent. We assume that the victim agent partially observes its environment and hence model it with a partially observable markov decision process (POMDP) (Åström, 1965). A POMDP generalizes the classic markov decision process to partial observability and is given by a tuple $\langle S, A, \mathcal{Z}, F, U, R, b_0 \rangle$, where $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{Z}$ are the state, action, and observation spaces,

respectively. Given a policy $\pi(a_t|o_{\leq t}, a_{<t})$, and starting from an initial belief state $b_0 : \mathcal{S} \to [0, 1]$, an agent chooses an action $a_t \in \mathcal{A}$ at time t where consequently the latent environment transitions to $s_{t+1} \sim F(s_{t+1}|s_t, a_t)$ upon which a noisy, occluded or otherwise incomplete observation $z_{t+1} \sim U(z_{t+1}|s_{t+1}, a_t)$ and a scalar reward $r_{t+1} \sim R(r_{t+1}|s_{t+1}, a_t)$ are returned to the agent. A commonly associated decision problem is one where the agent maximises the expected future return $\mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} r_t]$, given a discount factor where the expectation is taken over the trajectory of actions, states, and observed states and $0 \leq \gamma < 1$ is a discount factor.

POMDPs as belief MDPs. While POMDPs are not Markovian in observation space, agents can instead learn so-called *belief states*, which allows to transform the POMDP into a Markovian *belief MDP*. A belief MDP consists of belief states $b = \mathbb{P}(s_t|o_{\leq t}, a_{<t})$, where $\mathbb{P}$ denotes a probability distribution, and contains a *belief* transition function $\mathcal{K} : \mathcal{B} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$, as well as the belief-conditioned reward function $r : \mathcal{B} \times \mathcal{A} \to \mathbb{R}$. Here, $\mathcal{B}$ denotes the space of belief states $b : \mathcal{S} \to [0, 1]$.

Minimum-Norm Perturbation Attacks. In image classification tasks (Szegedy et al., 2013), an adversarial attack constructs a perturbation δ so that it causes a given classifier f to produce different predictions for x and $x + \delta$. The *minimum-norm perturbation (MNP)* adversarial attack seeks to find such perturbations δ that also satisfy budgeted constraints, *e.g.* $\|\delta\|_p \leq \epsilon$. It is important to note that MNP attacks can, in principle, attack a neural networks f in two ways: First, $x + \delta$ could be generated to perceptually resemble different inputs classes – which would not result in a perceptually invariant input – we call this type of attack a *semantic attack*. Second, $x + \delta$ could be semantically-invariant but still classified differently than the unperturbed input x; we call this type of attack a *neural attack*. Typically, the distance $\mathfrak{d}$ in observation space between samples belonging to different input types is much larger than the size of adversarial perturbations, i.e.

$$\mathfrak{d} = \min_{f(\mathbf{x}) \neq f(\mathbf{y})} \|\mathbf{x} - \mathbf{y}\|_p \gg \min_{(\mathbf{x}+\epsilon) \neq f(\mathbf{x})} \|\epsilon\|_p, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{D}_{\text{train}}.$$

That is to say, the smallest pixel distance between objects of different categories is typically much larger than the minimum perturbations required for neural attacks. Altogether, the MNP framework does not by itself allow to restrict attacks learnt to semantic attacks. Empirically, given typical perturbation budgets, MNP attacks are indeed found to heavily rely on *neural* attacks.

Minimum-Norm Perturbation Attacks on decision-making agents. Similarly, if an adversary has access to the observations of a decision-making victim agent with a neural network policy π_v, it can attack the victim by replacing its observations z_t by perturbed ones $\tilde{z}_t$. Such an attack is aimed at aligning the victim's behaviour with the adversary's arbitrary objective J_{adv}. The *minimum-norm perturbation (MNP)* adversary attack therefore seeks to identify and apply an additive perturbation $\epsilon \in \mathbb{R}^{\dim(\mathcal{Z})}$ to each agent observation z, where $\tilde{z} = z + \epsilon$, while minimising $\|\tilde{z} - z\|_p = \|\epsilon\|_p$ for some p-norm. Similarly to image classification, this can result in semantic or neural attacks, with the latter being much more frequent due to small perturbation budgets.

We define a scalar cost function for the adversarial agent, $R^{adv}\left(r_t^{adv}|s_t, a_t\right)$, note that the action is that taken by the victim agent. For the special case that $R^{adv} = -R$, the adversary's goal is exactly anti-aligned with the victim's goal. However, in general the adversary's goal does not need to be exactly anti-aligned with the victim's goal, but could be arbitrarily partially aligned.

We extend the classical formulation of minimum-norm perturbation attacks on MDPs (Kumar et al., 2021) to belief MDPs, with the adversaries objective therefore given as

$$\begin{aligned} J_{\text{adv}} &= \max_{\epsilon_0, \ldots \epsilon_{T-1}} \mathbb{E}_{\pi_v} \left[\sum_{t=0}^{T-1} \gamma^t R_{t+1}^{adv}(s_t, a_{t+1}) \right], \\ &\text{where } a_t \sim \pi_v\big(\cdot\,|\tilde{b}_t\big), s_{t+1} \sim F(\cdot|s_t, a_t), \\ &\text{s.t. } \|\epsilon_t\| \leq B \end{aligned} \tag{1}$$

Here $\tilde{b}_t$ denotes the victim's *perturbed belief state*, i.e. a belief based on a history of actions and *perturbed* observations.

4. Methods

4.1. Perceptual Incoherence under MNP attacks

We argue that *minimum-norm perturbation* attacks on POMDPs (see. Equation 1) may be detected by the victim agent, assuming that it has a model of the world in the form of a belief transition function $\mathcal{K}$. Given $\mathcal{K}$, the victim can, after each environment step, check whether the previous belief state transitions and the actions taken are permissible under $\mathcal{K}$.

More specifically, at time step t, the victim can compute the likelihood of the sequence of belief states, given the actions taken and the world model $\mathcal{K}$, i.e. $\mathbb{P}(\tilde{b}_{<t}, a_{<t}|\mathcal{K})$. The victim can use this to detect adversarial attacks if this likelihood is either zero or smaller than a threshold.

4.2. Illusionary Adversaries

To remain undetected, the adversary hence needs to fool $\mathcal{K}$ and π_v simultaneously with the same perturbation. We

assume that the victim agent computes the likelihood of the observed transitions as:

$$\begin{aligned}&\mathbb{P}(\tilde{b}_{<t}, a_{<t}|\mathcal{K}) \\ &= \mathcal{K}(\tilde{b}_{t-1}|a_{t-1}, \tilde{b}_{t-2})\pi(a_{t-1}|\tilde{b}_{t-2}) \cdot ... \cdot \mathbb{P}(\tilde{b}_0) \\ &= \mathbb{P}(\tilde{b}_0) \prod_{t'=1}^{t-1} \mathcal{K}(\tilde{b}_{t'}|a_{t'}, \tilde{b}_{t'-1})\pi(a_{t'}|\tilde{b}_{t'-1})\end{aligned} \tag{2}$$

Illusionary Attacks on Belief MDPs. We assume that the victim agent possesses a belief-state transition model $\mathcal{K}$ of the environment. We therefore define the objective of the illusionary attack as finding perturbations that maximize the objective function of the adversary while ensuring that the probability of observed belief-state transitions is larger than a threshold c, with $c > 0$:

$$\begin{aligned}&J_{\text{adv}} = \max_{\epsilon_0,...\epsilon_{T-1}} \mathbb{E}_{\pi_v}\left[\sum_{t=0}^{T-1} \gamma^t R^{adv}_{t+1}(s_t, a_{t+1})\right] \\ &\text{where } a_t \sim \pi_v\big(\cdot\,|\tilde{b}_t\big), s_{t+1} \sim F(\cdot|s_t, a_t), \\ &\text{s.t. } \mathbb{P}(\tilde{b}_{<t}, a_{<t}|\mathcal{K}) > c\end{aligned} \tag{3}$$

When is access to $\mathcal{K}$ realistic? Access to an environment $\mathcal{E}$ does not automatically grant access to $\mathcal{K} : \mathcal{B} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$, as $\mathcal{E}$ defines solely the *state transition function* $F : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, as well as the observation function U. Wherever $\mathcal{K}$ (or a suitable estimate $\hat{K}$) cannot be calculated from F, U in an analytically tractable fashion, one can attempt to approximate $\mathcal{K}$ through supervised training of a neural network $\mathcal{K}_\theta$. However, $\mathcal{K}_\theta$ might then itself be susceptible to MNP attacks, although these would require minimum-norm perturbations to be simultaneously feasible on both $\mathcal{K}_\theta$ and π_v, making such MNP attacks strictly harder. Instead, one could employ a non-neural pipeline using traditional scene understanding methods, such as fixed-feature object detectors, as an (approximation) to $\mathcal{K}$ that is not immediately susceptible to MNP attacks.

4.3. Fooling Humans.

We now investigate the special case in which, instead of a digitally represented belief transition model $\mathcal{K}$, victim agents employ a *human supervisor* in order to detect adversarial attacks at test time. This renders illusionary attack generation as in Equation 3 infeasible, as the volume of attainable human feedback is clearly highly limited.

While humans are not susceptible to neural attacks, they can be fooled by semantic attacks. Unlike neural attacks, semantic attacks are effectuated first in an abstract *entity space*, and only subsequently *rendered* to high-dimensional agent observations using access to the environment's observation function.

Entity POMDPs. To make the notion of semantic attacks more concrete, we introduce *Entity POMDPs* (Schroeder de Witt et al., 2019, cf. Dec-POMDP with Entities), in which each entity $e \in \mathcal{E}$ has a state representation s^e. At any time, the environment state is then given by $\mathbf{s}_t = \{s^e_t | e \in \mathcal{E}\} \cup \{\mathbf{s}^{\text{amb}}_t\} \in \mathcal{S}$, where $\mathbf{s}^{\text{amb}}_t$ subsumes any *ambient*, i.e. non-entity, state features. Each entity state feature contains two parts, $s^e = [f^e, \phi^e]$ with $\phi^e \in \Phi$ representing a (fixed) entity type, and f^e representing any other (dynamic) entity attributes (e.g. location, or velocity).

An explicit treatment of environment entities allows to reason over environment dynamics at a certain level of *semantic abstraction*, meaning we can express a subset of environment dynamics through discrete *rules* or *logical constraints*. For example, a common *physicality* constraint in environments of interest is that entities cannot change type over time, and that the dynamics of entity attributes are constrained by physics. As long an attacker respects these underlying *entity dynamics* $\mathfrak{D}$, and renders credible associated high-dimensional observations, humans are clearly unlikely to raise suspicion.

To guarantee MNP-free, **semantic illusionary adversarial attacks**, we follow Equation 3, but restrict the attacker's perturbation space to the entity space:

$$\begin{aligned}&J_{\text{adv}} = \max_{\epsilon_0,...\epsilon_{T-1}} \mathbb{E}_{\pi_v}\left[\sum_{t=0}^{T-1} \gamma^t R^{adv}_{t+1}(s_t, a_{t+1})\right] \\ &\text{where } a_t \sim \pi_v\big(\cdot\,|\tilde{b}_t\big), s_{t+1} \sim F(\cdot|s_t, a_t), \\ &\text{s.t. } \mathbb{P}(\tilde{b}_{<t}, a_{<t}|\mathfrak{D}) > c,\end{aligned} \tag{4}$$

Crucially, here the perturbed observations $\tilde{z} \neq z + \epsilon$, but rather $\tilde{z} = f(s^e + \epsilon, s^{\text{amb}})$, where f is the rendering function defined by the environment's observation function U.

Note that techniques introduced in this section may also be the attacker's preferred choice when it does not have access to (or is unsure about) $\mathcal{K}$ as used by the defender.

4.4. Realism feedback

While perfect illusionary attacks, i.e. those that fully comply with $\mathcal{K}$ (or $\mathfrak{D}$) may not be detectable, illusionary attacks that can only modify part of the agent's observation space may not always succeed. Hence, ensuring access to hardened *realism feedback channels* is a potentially powerful defense strategy.

5. Experimental Evaluation

5.1. Experimental Setup

The Environment and the Agent. We use a 2D gridworld (Minigrid, (Chevalier-Boisvert et al., 2018)) of variable size in which a single agent can navigate around by moving to

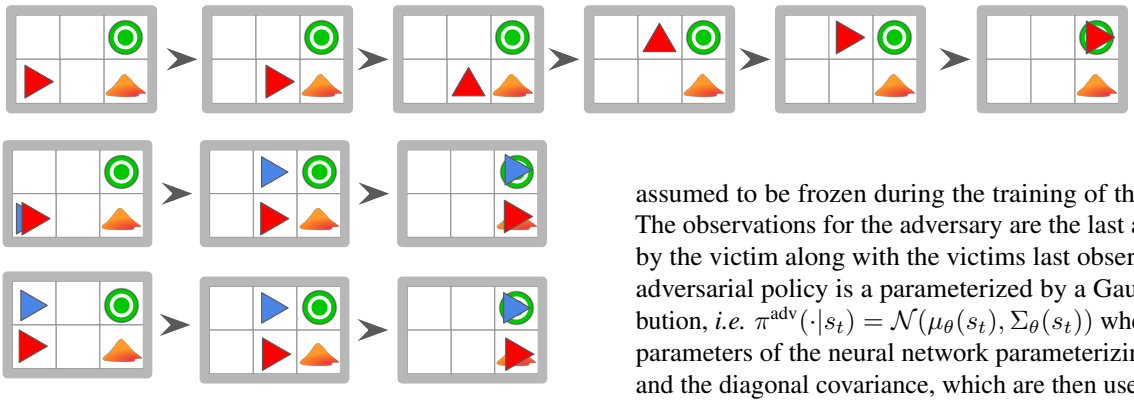

Figure 2: Environment that contains green goal cell (reward +1) and unfavourable cell with pile of red lava (reward -1). Top row: Trajectory of agent that executes optimal policy without perturbations. Middle row: Trajectory as observed by the victim agent under a minimum distance adversarial attack. Bottom row: Trajectory as observed by the agent under an illusionary adversarial attack. Perturbed observations in blue, true trajectories in red.

any of the four cells adjacent to the agent. The set of actions is given as $\mathcal{A} = \{\textit{move forward}, \textit{turn right}, \textit{turn left}\}$. The environment is deterministic, i.e. the state-transition probability of the environment is either 0 or 1, $T : \mathcal{S} \times A \rightarrow \{0, 1\}$. The lava environment in Figure 2 is fully observable (agent position and direction displayed by red triangle), while in the larger environment in Figure 4 the victim agent only partially observes the environment, indicated by the dashed cone (agent represented by camera). Throughout, we assume a fixed victim agent with a deterministic policy π_v trained until convergence using PPO (Schulman et al., 2017). We assume that the victim agent possesses a model of the world which accurately models the transition probabilities between different environment states, but observes no other variables except for its actions and observations; it does not observe the achieved reward which is a typical assumption for real-world deployment.

Adversary Agent. We now introduce our proposed non-embodied adversary agent. The adversary can modify the observations of the victim agent by taking actions on the environment as depicted in Figure 1. At each timestep, the adversary observes the previous state of the environment s_t, takes an adversarial action $a^{\text{adv}}_{t+1} \sim \pi^{\text{adv}}(\cdot|s_t)$ from a predefined set of actions $\mathcal{A}^{\text{adv}}$ and applies the action on the environment, generating the new perturbed observation $\tilde{z}_{t+1} = z_{t+1} + a^{\text{adv}}$ to be observed by the victim agent. The adversaries action space $\mathcal{A}^{\text{adv}}$ is different across the presented environments and defines the ability of the adversary to modify the different entities present in the environment. If the adversary takes an action that would yield an impossible state, the adversarial action is interpreted as a no action and thereof $\tilde{z}_{t+1} = z_t$. The policy of the victim agent is assumed to be frozen during the training of the adversary. The observations for the adversary are the last action taken by the victim along with the victims last observation. The adversarial policy is a parameterized by a Gaussian distribution, *i.e.* $\pi^{\text{adv}}(\cdot|s_t) = \mathcal{N}(\mu_\theta(s_t), \Sigma_\theta(s_t))$ where θ are the parameters of the neural network parameterizing the mean and the diagonal covariance, which are then used to sample the action of the adversary (which is afterward discretized). The reward function of the adversary is individually defined for the respective types of adversaries.

5.2. Adversaries in the Loop: Minimum Distance vs Illusionary

In this section, we show experimental results for rollouts of the victim agent under two adversaries, a MNP adversary and our proposed illusionary adversary. We consider the small environment shown in Figure 2. We assume that both adversaries are rewarded for minimizing the reward of the victim, *i.e.* , the goal of the adversary is to lure the victim agent into the lava cell at $(2, 3)$. For simplicity, we fix the initial state of the agent s_0 to cell $(2, 1)$, where the cell $(1, 1)$ is the top left most corner, facing to the right. Rolling out the victim agent under the adversary-free setting according to its policy π_v is shown in Figure 2 (top) where the agent correctly navigates to the green target cell. We note that, as this environment is fully observable, the belief states of the victim agent correspond to the actual environment state. We consider that the adversary performs (δ_x, δ_y) on the current position of the agent, i.e. $\mathcal{A}^{\text{adv}} = \{a^{\text{adv}} \in \mathbb{R}^2\}$ where $a^{\text{adv}} = [\delta_x, \delta_y]^\top$. That is to say, if the adversarial action is $a^{\text{adv}} = [0, 1]^\top$ for an agent at position $(2, 1)$, then the agent observes its perturbed state as $(1, 1)$.

Minimum Distance Adversary. Here, the adversary receives as reward the negative of the victim agents reward, and an additional negative reward corresponding to the norm of each induced perturbation. An optimal minimum-norm perturbation attack is displayed in Figure 2 (middle row). Only the observation of the victim agent at cell $(2, 2)$ is perturbed. In this setting, the observed state $\tilde{s}_t$, compared to the true environment state s_t, differs only by one one value. Such adversary, as shown in the middle row of Figure 2, will correctly lure the agent into the lava cell. It is easy to observe that such adversary is not consistent with the state-transition probability of the environment, *i.e.* , it is not world-model aligned, which can be observed in the Figure as the agent "jumps" between cells $(2, 1)$ and $(1, 2)$, irrespective of having taken the $forward$ action.

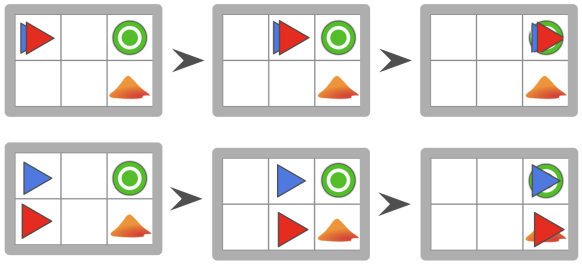

Figure 3: (True states in red, observed states in blue.) This Figure shows the perturbation added by the illusionary adversary for two different starting positions of the victim agent. If the agent starts in the top left cell, no perturbation is added by the illusionary adversary as the victim agent cannot be lured into the bottom right cell without violating the agent's world model. If the agent starts in the bottom left cell, the adversary chooses the correct illusionary perturbations to lure the agent into the lava cell at the bottom right.

Illusionary Adversary. Here, the adversary receives as reward the negative of the victim agents reward, and an additional negative reward for actions that result in observations that are impossible under the victim agent's belief (state) transition model $\mathcal{K}$. Unlike minimum distance attacks, our proposed illusionary adversary that maximizes the objective in equation 3 is world-model aligned. It correctly lures the victim agent to the lava cell just like the minimum distance adversary. However, the illusionary adversary perturbs the observed position of the victim agent consistently, as it can be observed in Figure 2 (bottom row). This simple experiment supports our hypothesis that MNP attacks may be easily detected by agents with a world model.

Worldmodel Aligned Adversaries trade off reward for remaining undetected. Consider the previously introduced environment depicted in Figure 3. We now modify this environment, such that the victim agent's initial position is randomly chosen to be either $(1, 1)$ or $(2, 1)$. We observe in Figure 3 that the adversary agent trades off achieved reward for remaining undetected, i.e. ensuring that all observations are worldmodel-aligned.

Detectable Adversarial Attacks. We now consider the environment depicted in Figure 4, where the victim agent has to navigate to the green goal position (which it does not observe), by using the red and blue square as landmarks. We significantly limit the influence that the adversary has on the victim agent's observation. Instead of being able to completely change the observation of the agent by modifying the position from which it observes the world, the adversary can now only modify the positions of the two red and blue landmark tiles (entities). Thereby, the agent could detect the adversarial attack, as the adversary does not have access to the victim's full observation, but only to the two landmarks.

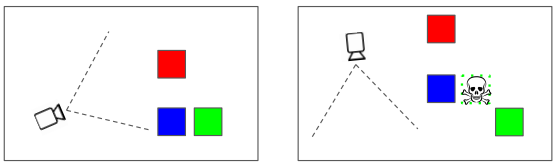

Figure 4: This large environment awards the agent with +1 for reaching the green goal square; no other rewards exist. The agents field of view is indicated by the camera viewing angles. The agent does not observe the goal square any different from other squares and therefore relies on the landmarks (red and blue colored cells) for navigating to the goal. The adversary agent can perturb the observation of the agent by shifting the landmarks. Left: The true environment state. Right: The perturbed environment state as observed by the agent.

The adversary hence chooses a perturbation action at each time step which modifies the observed position of both the blue and red landmark by $[-1, 0, +1]$ in x and y individually. The reward of the adversary agent is further changed such that it performs a targeted attack, by rewarding it for luring the victim agent into the cell that is northwest of the actual goal position (indicated by the skull in Figure 4). We again pretrain the victim agent and afterward train the adversary until convergence. We find that the adversary has learned to perform an illusionary attack on the victim agent by shifting the positions of both landmarks in northwest direction (see Figure 4, right image).

However, in this environment the victim agent could have detected the adversary by having a policy that seeks *realism feedback*. The victim agent could detect the adversary by using its world model to find that the distance between the landmarks and the walls has changed. It could then have developed a robust policy that relies on the walls instead of the landmarks, and could thereby successfully navigate to the true goal position under adversarial perturbations.

6. Discussion and Conclusion

In this paper, we introduce *illusionary attacks*, a novel form of adversarial attacks on sequential decision makers, that can fool even agents with access to transition models. Importantly, we show that illusionary attacks, unlike traditional minimum-norm perturbation adversarial attacks, can fool human supervisors. We show how defenders may employ realism feedback in order to mitigate the chance of being fooled by an illusionary attack. This implies that any real-world autonomous decision making systems need to provision for adequate, hardened realism feedback channels, and maximise their utility through test time policies that employ information gathering and environment probing functionality.

Acknowledgements

This work is supported by the UKRI grant: Turing Al Fellowship EP/W002981/1 and EPSRC/MURI grant: EP/N019474/1. CS is generously sponsored by the Cooperative AI Foundation.

References

Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations. 2020.

Andrychowicz, O. M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 2020.

Behzadan, V. and Munir, A. Whatever does not kill deep reinforcement learning, makes it stronger. *arXiv preprint arXiv:1712.09344*, 2017.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Chaubey, A., Agrawal, N., Barnwal, K., Guliani, K. K., and Mehta, P. Universal Adversarial Perturbations: A Survey. Technical report, May 2020.

Chen, T., Liu, J., Xiang, Y., Niu, W., Tong, E., and Han, Z. Adversarial attack and defense in reinforcement learning-from AI security view. *Cybersecurity*, 2019.

Chevalier-Boisvert, M., Willems, L., and Pal, S. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid, 2018.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*. PMLR, 2019.

Everett, M., Lutjens, B., and How, J. P. Certifiable Robustness to Adversarial State Uncertainty in Deep Reinforcement Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Gleave, A., Dennis, M., Wild, C., Kant, N., Levine, S., and Russell, S. Adversarial Policies: Attacking Deep Reinforcement Learning. Technical report, arXiv, 2021.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR, 2018.

Havens, A., Jiang, Z., and Sarkar, S. Online robust policy learning in the presence of unknown adversaries. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

Hussenot, L., Geist, M., and Pietquin, O. Copycat: Taking control of neural policies with constant attacks. *arXiv preprint arXiv:1905.12282*, 2019.

Kos, J. and Song, D. Delving into adversarial attacks on deep policies. *arXiv preprint arXiv:1705.06452*, 2017.

Kumar, A., Levine, A., and Feizi, S. Policy Smoothing for Provably Robust Reinforcement Learning. Technical report, arXiv, 2021.

Lin, Y.-C., Hong, Z.-W., Liao, Y.-H., Shih, M.-L., Liu, M.-Y., and Sun, M. Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. 2017.

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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 2015.

Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., and Chowdhary, G. Robust Deep Reinforcement Learning with Adversarial Attacks. In *AAMAS*, 2018.

Qiaoben, Y., Ying, C., Zhou, X., Su, H., Zhu, J., and Zhang, B. Understanding Adversarial Attacks on Observations in Deep Reinforcement Learning. Technical report, arXiv, 2021.

Salimans, T., Ho, J., Chen, X., Sidor, S., and Sutskever, I. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Schroeder de Witt, C., Foerster, J., Farquhar, G., Torr, P., Boehmer, W., and Whiteson, S. Multi-Agent Common Knowledge Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2019.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms. Technical report, arXiv, August 2017.

Sun, J., Zhang, T., Xie, X., Ma, L., Zheng, Y., Chen, K., and Liu, Y. Stealthy and Efficient Adversarial Attacks against Deep Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tekgul, B. G., Wang, S., Marchal, S., and Asokan, N. Real-time attacks against deep reinforcement learning policies. *arXiv preprint arXiv:2106.08746*, 2021.

Todorov, E., Erez, T., and Tassa, Y. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, October 2012.

Wu, F., Li, L., Huang, Z., Vorobeychik, Y., Zhao, D., and Li, B. CROP: Certifying Robust Policies for Reinforcement Learning through Functional Smoothing. *ArXiv*, 2021.

Zintgraf, L. M., Shiarlis, K., Igl, M., Schulze, S., Gal, Y., Hofmann, K., and Whiteson, S. VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning. *International Conference on Learning Representations*, 2020.

Åström, K. J. Optimal Control of Markov Processes with Incomplete State Information I. *Journal of Mathematical Analysis and Applications*, 1965.