Illusory Attacks: Detectability Matters in Adversarial Attacks on Sequential Decision-Makers

Anonymous Authors[1]

Abstract

Autonomous agents deployed in the real world need to be robust against adversarial attacks on sensory inputs. Robustifying agent policies requires anticipating the strongest attacks possible. We demonstrate that existing observation-space attacks on reinforcement learning agents have a common weakness: while effective, their lack of temporal consistency makes them *detectable* using automated means or human inspection. Detectability is undesirable to adversaries as it may trigger security escalations. We introduce *perfect illusory attacks*, a novel form of adversarial attack on sequential decision-makers that is both effective and provably *statistically undetectable*. We then propose the more versatile R-illusory attacks, which result in observation transitions that are consistent with the state-transition function of the adversary-free environment and can be learned end-to-end. Compared to existing attacks, we empirically find R-illusory attacks to be significantly harder to detect with automated methods, and a small study with human subjects[1] suggests they are similarly harder to detect for humans. We propose that undetectability should be a central concern in the study of adversarial attacks on mixed-autonomy settings.

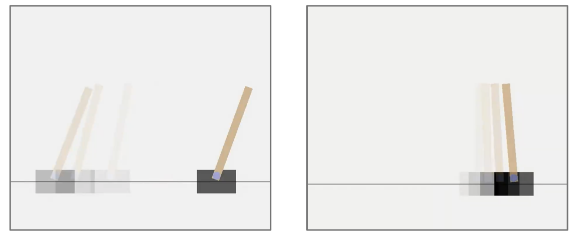

Figure 1: The left image shows an observation sequence (older observations are faded out) as seen under a state-of-the-art adversarial attack. One easily classifies this sequence as *adversarially attacked* as the cart appears to jump horizontally, violating the transition dynamics of the unattacked system. The sequence in the image on the right, resulting from our proposed *illusory attacks*, appears unsuspicious.

1. Introduction

Deep reinforcement learning algorithms [52, 66, 25, 64] have been applied to numerous sequential decision-making problems, ranging from recreational games [3], to robotics [79, 2], nuclear fusion [17], and solar geoengineering [16]. In autonomous driving, deep neural networks are increasingly used for vision-related control tasks involving object detection [61, 60, 92, 88, 42, 68] and segmentation [27, 51], lane detection [7, 55], or depth estimation [83]. However, the susceptibility of deep neural networks to adversarial attacks poses threats to their safety-critical application [38, 30]. This motivates research into strong adversarial attacks and robustification against them [90, 72, 47].

Autonomous AI systems deployed to the real world often feature a combination of both automated and human security monitoring [82]. In practice, strong attackers seek to evade detection as attacked agents might have access to contingency options such as executing an emergency shutdown or triggering security escalations [11]. A prime example of such behaviour are security incidents where attackers feed unsuspicious pre-recorded input to surveillance cameras, or an industrial control panel [43, STUXNET 417 attack]. We argue that attack detectability should become a central consideration when safeguarding the robustness of (Human-)AI systems to adversarial attacks.

Existing frameworks for observation-space adversarial attacks on sequential-decision makers are often not primarily concerned with detectability [90, 72], be it by automated means, or by human inspection. In this paper, we introduce the *illusory attack framework*. Unlike most existing types of observation-space adversarial attacks, illusory attacks are not fundamentally constrained by perturbation budgets, but detectability concerns.

To illustrate the need for our new attack framework, we

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Under review by *AdvML-Frontiers'23: The 2nd Workshop on New Frontiers in Adversarial Machine Learning@ICML'23*, Honolulu, Hawaii, USA. Do not distribute.

[1]IRB approval under reference xxxxxx/xxxxx

first construct simple automated detectors and empirically show that these can detect state-of-the-art observation-space attacks with high probability across a variety of simulated environments. Such automated detectors merely require victim agents to have access to a (possibly approximate) model of the state-transition function. Access to such *world models* is a common assumption in autonomous systems literature [24, 74], *e.g. world models* can be learned from train-time experience [73, DYNA]. Likewise, we show that humans can detect state-of-the-art observation-space attacks through visual inspection (see Figure 1 for an illustration).

We then show that our novel framework gives rise to adversarial attacks that are harder to detect, or indeed statistically undetectable, by both automated means and humans. To illustrate the latter, we construct *perfect illusory attacks*, a novel class of *provably statistically undetectable* adversarial attacks, and implement these in various standard benchmark environments. However, we prove that there are environments that do not admit *perfect* illusory attacks. We thus introduce the more versatile R-illusory attacks attacks, which are a relaxation of perfect illusory attacks that, while being statistically detectable in theory, can be learned through end-to-end gradient-based optimisation and can dynamically trade-off between detectability and adversarial performance.

We empirically confirm that R-illusory attacks can be efficiently learned, and are much harder to detect than existing observation-space adversarial attacks - both using automated detectors based on world models, as well as for humans. We also find that that existing robustification methods are largely ineffective against R-illusory attacks in practice. This suggests that the implementation and effective use of *reality feedback channels*, i.e., observation channels that are hardened against adversarial interference, will be of fundamental importance in the quest to adversarially robustify real-world mixed- and shared-autonomy, and (Human)-AI systems.

Our work makes the following contributions:

- We demonstrate that state-of-the-art adversarial attacks are reliably detected both with simple automated detectors, as well as by human inspection.
- We formalise the novel *illusory* attack framework and show that it gives rise to *perfect illusory attacks*, which are statistically undetectable observation-space adversarial attacks (see Section 3.5).
- We introduce R-illusory attacks, a relaxation of perfect illusory attacks that can be learned using gradient-based optimisation (see Section 3.8) and, as we show empirically, are significantly harder to detect by both automated means and human inspection.

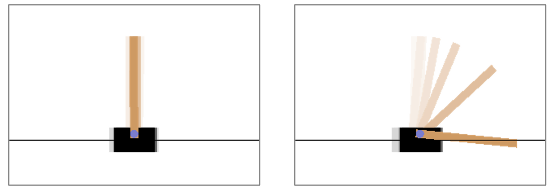

Figure 2: In *CartPole*, the agent aims to balance the brown pole by adjusting the position of the black cart. In the perfect illusory attack depicted above, the agents observations (left) appear unperturbed while the true system fails (right).

2. Background and notation

We denote a probability distribution over a set $\mathcal{X}$ as $\mathcal{P}(\mathcal{X})$, and an unnamed probability distribution as $\mathbb{P}(\cdot)$. The empty set is denoted by $\emptyset$, and the unit impulse as $\delta(\cdot)$.

MDP and POMDP. A Markov decision process (MDP) [6] is a tuple $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$, where $\mathcal{S}$ is the finite[2] non-empty state space, $\mathcal{A}$ is the finite non-empty action space, $p : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{P}(\mathcal{S})$ is the probabilistic state-transition function, and $r : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{P}(\mathbb{R})$ is a bounded reward function, *i.e.* $\forall (s, a) \in \mathcal{S} \times \mathcal{A},\ |r(s, a)| \leq \mathcal{R}$ almost surely for some finite $\mathcal{R} > 0$. Starting from a state $s_t \in \mathcal{S}$ at time t, an action $a_t \in \mathcal{A}$ taken by the agent policy $\pi : \mathcal{S} \mapsto \mathcal{P}(\mathcal{A})$ effects a transition to state $s_{t+1} \sim p(\cdot|a_t)$ and the emission of a reward $r_{t+1} \sim r(\cdot|s_{t+1}, a_t)$. We define the initial system state at time $t = 0$ is drawn as $s_0 \sim p(\cdot|\emptyset)$. For simplicity, we consider episodes of infinite horizon and hence introduce a discount factor $0 \leq \gamma < 1$. In a partially observable MDP [94, 35, POMDP] $\langle \mathcal{S}, \mathcal{A}, \Omega, \mathcal{O}, p, r, \gamma \rangle$, the agent does not directly observe the system state s_t but instead receives an observation $o_t \sim \mathcal{O}(\cdot|s_t)$ where $\mathcal{O} : \mathcal{S} \mapsto \mathcal{P}(\Omega)$ is an observation function and Ω is a finite non-empty observation space. The canonical embedding *pomdp* : $\mathfrak{M} \hookrightarrow \mathfrak{P}$ from the set of finite MDPs $\mathfrak{M}$ to the family of POMDPs $\mathfrak{P}$ maps $\Omega \mapsto \mathcal{S}$, and sets $\mathcal{O}(s) = s,\ \forall s \in \mathcal{S}$. In a POMDP, the agent acts on a policy $\pi : \mathcal{H}^*_{\setminus r} \mapsto \mathcal{P}(\mathcal{A})$, growing a history $h_{t+1} = h_t a_t o_{t+1} r_{t+1}$ from a set of histories $\mathcal{H}^t := (\mathcal{A} \times \mathcal{O} \times \mathbb{R})^t$, where $\mathcal{H}^* := \bigcup_t \mathcal{H}^t$ denotes the set of all finite histories. We denote histories (or sets of histories) from which reward signals have been removed as $(\cdot)_{\setminus r}$, and the distribution over a history h_t as $\mathbb{P}^\pi_t$. In line with standard literature [53], we distinguish between two stochastic processes that are induced by pairing a POMDP with a policy π: The *core process*, which is the process over state random variables $\{S_t\}$, and the *observation process*, which is induced by observation random variables $\{O_t\}$. We also define the *reward process*

[2]For conciseness, we restrict our exposition to finite state, action and observation spaces. Results carry over to continuous state-action-observation spaces under some technical conditions that we omit for brevity [76].

over reward random variables as $\{R_t\}$, where $t \in \mathbb{N}_0$. The frequentist agent's goal is then to find an optimal policy π^* that maximises the total expected discounted return, i.e. $\pi^* = \arg\sup_{\pi\in\Pi} \mathbb{E}_{h_\infty \sim \mathbb{P}^\pi_\infty} \sum_{t=0}^{\infty} \gamma^t r_t$, where $\Pi := \{\pi : \mathcal{H}^*_{\setminus r} \mapsto \mathcal{P}(\mathcal{A})\}$ is the set of all policies.

Observation-space adversarial attacks. Observation-space adversarial attacks consider the scenario where an *adversary* manipulates the observation of a *victim* at test-time. Much prior work falls within the SA-MDP framework [90], in which a state-adversarial agent with policy $\xi : \mathcal{S} \mapsto \mathcal{P}(\mathcal{S})$ generates adversarial observations $o_t \sim \xi(s_t)$. The perturbation is bounded by a budget $\mathcal{B} : \mathcal{S} \mapsto 2^{\mathcal{S}}$, limiting supp $\xi(\cdot|s) \in \mathcal{B}(s)$. For simplicity, we consider only zero-sum adversarial attacks, where the adversary minimizes the expected return of the victim. In case of *additive* perturbations $\epsilon_t \in \mathcal{S}$ [40], $\xi(s_t) := \delta(o_t)$. Here, $o_t := s_t + \epsilon_t$, and o_t are subject to a real positive episodic perturbation budget $B \in \mathbb{R}$. Given a victim policy π_v, this yields the following definition of an *optimal state-conditioned observation-space adversary*:

$$\xi^* = \arg\min_{\xi} \mathbb{E}_{\pi_v}[\textstyle\sum_{t=0}^{\infty} r_t] \qquad \text{s. t. } \textstyle\sum_{t=0}^{\infty} \|\epsilon_t\|_2^2 \leq B^2, \tag{1}$$

where $a_t \sim \pi_v(\,\cdot\,|o_t), o_t \sim \xi(s_t), s_{t+1} \sim p(\cdot|s_t, a_t)$. In other words, given a fixed victim policy, the adversary seeks perturbations that minimise the victim return under the given budget constraints.

3. Illusory attacks

3.1. The illusory attack framework

We introduce a novel *illusory* attack framework in which an adversary and a victim share the same environment $\mathcal{E}$ at test time, thus inducing a two-player zero-sum game $\mathcal{G}$ [81]. The following facts about $\mathcal{G}$ are commonly known [26] between adversary and victim: At test time, the adversary performs observation-space attacks on the victim. The victim can sample from the environment shared with an arbitrary adversary at train time, but has no certainty over which specific test-time policy the adversary will choose. The adversary can sample from the environment shared with an arbitrary victim at train time, but has no certainty over which specific test-time policy the victim will choose. The task of the victim is to act optimally with respect to its expected test-time return, while the task of the adversary is to minimise the victim's expected test-time return. Unlike in prior work (see Section 2), the adversary's observation-space attacks are not necessarily bounded by perturbation budgets.

We assume that the victim's reward signal is endogenous [4], which means it depends on the victim's action-observation history and is not explicitly modeled at test-time. This exposes the victim's test-time reward signal to manipulation by the adversary. We note that even if it is assumed that the victim's reward signal is supplied exogenously, reinforcement learning environments of interest frequently emit sparse or delayed reward signals that are rather uninformative about the current environment state.

Definition 3.1 (Test-time decision process)**.** We denote the stochastic process induced by sharing an environment $\mathcal{E}$ between a victim with policy π_v and an adversary with policy ν as $\mathcal{E}^{\pi_v}_\nu$ For simplicity, we assume that $\mathcal{E}^{\pi_v}$, i.e. the special case in which the adversary chooses a policy that leaves the victim's observations $o_t \in \Omega$ unaffected, reduces to π_v acting in a finite MDP $\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$ with infinite horizon (see Section 2). We assume that both the victim policy $\pi_v : \mathcal{H}^*_{\setminus r} \mapsto \mathcal{P}(\mathcal{A})$ and the state-observing adversary policy $\nu : \mathcal{S} \times \mathcal{H}^*_{\setminus r} \mapsto \mathcal{P}(\Omega)$ are history-dependent. The semantics of $\mathcal{E}^{\pi}_\nu$ are as follows: At time $t = 0$, we sample an initial state $s_0 \sim p(\cdot|\emptyset)$. The adversary then samples an observation $o_0 \sim \nu(\cdot|s_0)$ which is emitted to the victim. The victim takes an action $a_0 \sim \pi(\cdot|o_0)$, upon which the state transitions to $s_1 \sim p(\cdot|s_0, a_0)$. At time $t > 0$, the victim has accumulated a history $h_t := o_0 a_0 r_1 \ldots, o_{t-1}$ upon which $o_t \sim \nu(\cdot|s_t, h_{t\setminus r})$ conditions.

We are ultimately interested in characterising the Nash equilibria induced by $\mathcal{G}$ [54]. To this end, we now show that, for any choice of ν , the victim's task of finding an optimal policy in $\mathcal{E}^{(\cdot)}_\nu$ is equivalent to instead finding an optimal policy in a corresponding POMDP $\mathcal{E}_e\,(\mathcal{E}^{(\cdot)}_\nu)$.

Theorem 3.2 (POMDP correspondence)**.** *For any $\mathcal{E}^{(\cdot)}_\nu$, there exists a corresponding POMDP $\mathcal{E}_e\,(\mathcal{E}^{(\cdot)}_\nu)$ for which the victim's learning problem is identical. See Appendix B.1 for a proof.*

Theorem 3.2 implies that, given enough memory [89] , the adversary can be chosen such that the state-space of $\mathcal{E}_e(\mathcal{E}^{(\cdot)}_\nu)$ becomes arbitrarily due to its infinite horizon. This renders the worst-case problem of finding an optimal victim policy in $\mathcal{E}_e(\mathcal{E}^{(\cdot)}_\nu)$ intractable even if the adversary's policy is known [33, 45]. The underlying game $\mathcal{G}$, therefore, assumes an infinite state space, preventing recent progress in solving finite-horizon extensive-form games [39, 50, 69] from being leveraged in characterising its Nash equilibria. For the remainder of this paper, we overload $\mathcal{E}^{(\cdot)}_\nu$ to instead refer to its corresponding POMDP.

3.2. Adversary detection as a defense strategy

We investigate a victim strategy that, instead of a priori finding the best response to an unknown adversary, focuses on efficiently detecting whether an effective test-time adversary is present, *i.e.* whether the victim is deployed to $\mathcal{E}^{\pi_v}_\nu$ or $\mathcal{E}^{\pi_v}$. Knowing whether an effective adversary is present in the environment would allow a real-world agent to take contingency options that would suspend operations or trigger security escalations[3].

To understand the limits of adversary detection that the victim can achieve, it is important to recall the concept of *stochastic equivalence* between stochastic processes.

Definition 3.3 (Stochastic equivalence)**.** Two stochastic processes $X_1(t)$ and $X_2(t),\ t \in T$, defined on a common probability space are called stochastically equivalent if for any $t \in T$, $\mathbb{P}(X_1(t) = X_2(t)) = 1$, i.e. if the random variables at each time-step are almost surely following the same distribution. [36, 19]

Theorem 3.4 (The limits of adversary detection)**.** *At test-time, the victim only has access to observation samples $o_t \in \Omega$ from $\mathcal{E}_\nu^{\pi_v}$. The detection task is thus limited to establishing equivalence [62] of the observation processes induced by $\mathcal{E}_\nu^{\pi_v}$ and $\mathcal{E}^{\pi_v}$ (see Section 2). As $\mathcal{E}^{\pi_v} := pomdp(\mathcal{M})$, where $\mathcal{M} := \langle \Omega, \mathcal{A}, p, \cdot, \cdot \rangle$ is a finite MDP, the detection task can be decomposed into a sequence of empirical hypothesis tests:*

1. *Test whether $\Omega = \mathcal{S}$ [Observation space compatibility].*
2. *Test whether $\mathbb{P}(o_{t+1}|o_t, a_t) = \mathcal{O}'(o_{t+1}|o_t, a_t)$ [State transition equality].*
3. *Test whether $\mathbb{P}(o_{t+1}|o_{\leq t}, a_{\leq t}) = \mathbb{P}(o_{t+1}|o_t, a_t)$ [Conditional independence[4]].*

Here, $\mathcal{O}'$ is the observation function of $\mathcal{E}^{\pi_v}$. Note that these tests could be conducted either by comparing cross-sectional distributions or autocorrelations. If any of these tests fails, then the victim does not act in $\mathcal{E}^{\pi_v}$. If all tests pass, then the observation processes of $\mathcal{E}_\nu^{\pi_v}$ and $\mathcal{E}^{\pi_v}$ are stochastically equivalent (see Def. 3.3). However, this does not establish the equivalence of the underlying core and reward processes. The proof follows by definition (see Appendix B.2).

Note that following the protocol of Theorem 3.4 requires the victim to *sample estimates* of the distributional quantities involved. In practice, the victim's ability to detect the presence of an effective adversary is therefore also constrained by the number of attainable test-time samples.

3.3. Illusory attacks

We now concern ourselves with characterising the best responses that the adversary can choose when facing a victim agent pursuing the detection protocol defined in Theorem 3.4. We assume that the adversary cannot fully characterise the victim's contingency options. Hence, detection poses an unquantifiable risk to the adversary, which we assume it prioritises to avoid.

[3] Modeling such contingency options as part of an extended game $\mathcal{G}'$ lies outside the scope of this paper.

[4] This is to exclude long-term correlations, see Appendix B.4.

3.3.1. Perfect illusory attacks

Theorem 3.5 (Existence of perfect illusory attacks)**.** *Given $\mathcal{E}_{(\cdot)}^{\pi_v}$, the adversary can sometimes choose a perfect illusory attack ν such that, simultaneously,*

- *The core and reward processes of $\mathcal{E}_\nu^{\pi_v}$ and $\mathcal{E}^{\pi_v}$ differ.*
- *The victim cannot distinguish $\mathcal{E}_\nu^{\pi_v}$ from $\mathcal{E}^{\pi_v}$ using the protocol defined in Theorem 3.4 even when given an unlimited number of test-time samples.*

For a proof see Appendix B.3.

We provide examples of perfect illusory attacks in Section 4 as well as in the results video in the supplementary material. Clearly, in addition to remaining undetectable by the victim, the adversary should choose an attack from the set of perfect illusory attacks that minimise the victim's expected test-time return to the greatest possible extent.

Definition 3.6 (Optimal illusory attack)**.** An optimal illusory attack ν^* on $\mathcal{E}_{(\cdot)}^{\pi_v}$ is the subset of perfect illusory attacks $\{\nu\}$ corresponding to the highest expected adversarial test-time return, *i.e.* ,

$$\nu^* = \arg\inf_{\nu} \mathbb{E}_{h\sim\mathcal{E}_\nu^{\pi_v}}\left[\textstyle\sum_{t=0}^{\infty} r_t\right], \text{s.t. } \mathbb{P}_{\mathcal{E}_\nu^{\pi_v}} = \mathbb{P}_{\mathcal{E}^{\pi_v}}. \quad (2)$$

Here, we use the shorthand $\mathbb{P}_{\mathcal{E}_\nu^{\pi_v}} = \mathbb{P}_{\mathcal{E}^{\pi_v}}$ to imply that, $\forall h \in \mathcal{H}_{\setminus r}^*$, the probability distributions over h are equal for both processes $\mathcal{E}_\nu^{\pi_v}$ and $\mathcal{E}^{\pi_v}$.

3.3.2. n-step perfect illusory attacks

Perfect illusory attacks are not always possible to construct (see Appendix B.3). To arrive at a practical relaxation of perfect illusory attacks, we note that testing for conditional independence of observation transitions (Step 3 in the victim detection protocol defined in Theorem 3.4) can require a large amount of test-time samples [67], making it difficult to achieve for long time windows in many practical settings. Hence we define *n-step perfect illusory attacks*, which only preserve conditional independence of observation transitions over up to n time steps.

Definition 3.7 (n-step perfect illusory attacks)**.** Given $\mathcal{E}_{(\cdot)}^{\pi_v}$, the set of n-step perfect illusory attacks is given by all ν for which the *observation process* of $\mathcal{E}_\nu^{\pi_v}$ satisfies that $\mathbb{P}(o_t|o_{<t}, a_{<t}) = \mathbb{P}(o_t|o_{t-1,\ldots,t-n}, a_{t-1,\ldots,t-n}),\ \forall t \geq 0$.

We omit the analogous definition of *n-step optimal illusory attacks* for brevity.

3.3.3. R-illusory attacks

As we wish to perform gradient-based optimisation of illusory attacks, we in practice approximate the conditional independence constraints using a Lagrangian relaxation. The resulting R-illusory attacks are an approximation of 1-*step*

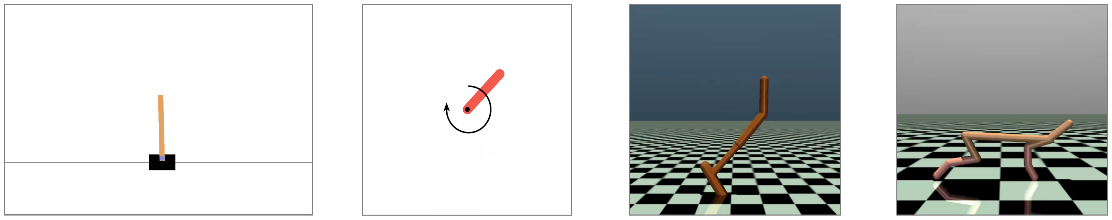

Figure 3: Benchmark environments used for empirical evaluation, from left to right. In *CartPole*, the agent has to balance a pole by moving the black cart. In *Pendulum*, the agent has to apply a torque action to balance the pendulum upright. In *Hopper* and *HalfCheetah*, the agent has to choose high-dimensional control inputs such that the agent moves towards the right of the image.

Attack	Normalised adversary score in % no defence	smoothing	ATLA	ATLA abl.
MNP [40]	96 ± 4	95 ± 1	-	-
SA-MDP [91]	98 ± 7	68 ± 3	89 ± 6	93 ± 7
R-illusory attack (ours)	77 ± 7	65 ± 4	72 ± 4	72 ± 6
Perfect illusory attack (ours)	63 ± 8	71 ± 1	73 ± 7	73 ± 5

Table 1: Adversary scores and standard deviations averaged across environments for different defence methods and different attacks ($\beta = 0.2$). Defences decrease the adversary score, i.e., increase the victim reward across all classes and all attack algorithms. We find that most defences only lead to marginal changes (see App. B.5 for discussion.)

optimal illusory attacks that weighs off 1-step detectability and adversary task performance.

Definition 3.8 (R-illusory attacks)**.** R-illusory attacks are a Lagrangian relaxation of 1-step optimal illusory attacks:

$$\nu_R = \arg\inf_{\nu} \mathbb{E}_{h\sim\mathcal{E}_{\nu}^{\pi_v}} \sum_{t=0}^{\infty} r_t + \lambda \mathcal{D}\left[\mathbb{P}_{\mathcal{E}_{\nu}^{\pi_v}}(\cdot|o_t, a_t), p(\cdot|o_t, a_t)\right], \tag{3}$$

where $\lambda > 0$ is a hyper-parameter that determines the weighing of the two objectives, and $\mathcal{D}$ is a distance measure between probability distributions.

4. Empirical evaluation of illusory attacks

We now compare and contrast perfect illusory attacks and R-illusory attacks with state-of-the-art observation-space adversarial attacks according to the criteria of detectability, adversarial performance, and attainable robustification. To this end, we first construct a simple adversarial attack detector Δ that implements a CUSUM-like decision rule [57] based on a learned model of the observation-transition function of the adversary-free environment. We provide the source code, a summary video, and individual videos per attack and random seed in the supplementary material.

Experimental setup. We evaluate our methods on four standard benchmark environments (see Figure 3) with continuous state spaces whose dimensionalities range from 1 to 17 and both continuous and discrete action spaces. The mean and standard deviations of both detection and performance results are estimated from 200 independent episodes per each of 5 random seeds. Victim policies are pre-trained in unattacked environments, and frozen during adversary training. We assume the adversary has access to the unattacked environment's state-transition function t. For implementation details please refer to the appendix.

Empirical detector evaluation. We now compare and contrast perfect illusory attacks, R-illusory attacks, MNP attacks and SA-MDP attacks across different perturbation budgets β. We found results for $\beta = 0.05$ to be comparable to those for $\beta = 0.2$ and focus on $\beta = 0.2$ in this analysis (see appendix for all results). As detailed in Figure 5, the detector Δ detects MNP and SA-MDP attacks with a probability close to 100%. In contrast, the detector classifies R-illusory attacks as adversarially attacked with very low probability. Note that perfect illusory attacks, as per definition, are undetectable and do not have a perturbation budget. In Figure1, we illustrate why SA-MDP attacks can be detected with world models.

Detection by human inspection. We investigate whether humans unfamiliar with adversarial attacks can detect adversarial attacks in simple and easy-to-understand environments. We omit perfect illusory attacks in the study as these, by definition, cannot be detected. We found that participants classified MNP and SA-MDP attacks as suspicious with high accuracy (see *Human detection* in Figure 5). In contrast, participants were almost equally likely to classify videos of unattacked and R-illusory attacked trajectories as unsuspicious. In fact, at a confidence level of 95%, the hypothesis *participants are equally likely to classify an unattacked sequence as attacked as to classify an R-illusory attacked sequence as attacked* cannot be rejected. Our findings suggest that humans are unable to detect R-illusory attacks from short observation sequences in our simple environments. See Appendix B.11 for full results and the corresponding z-test statistic.

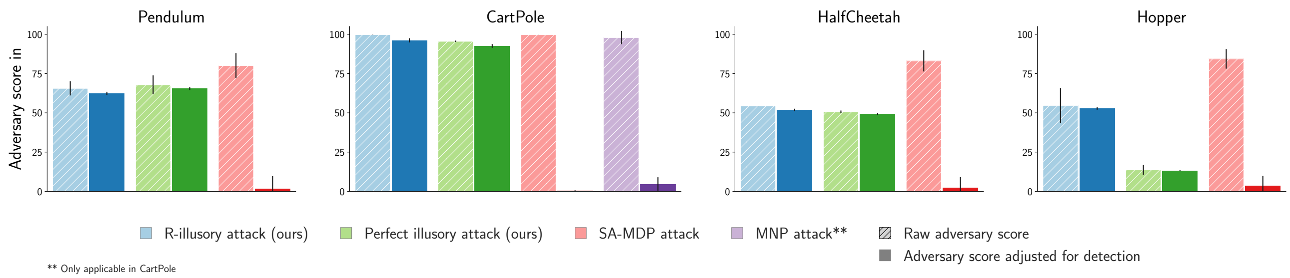

Figure 4: We display normalised adversary scores, indicating the reduction in the victim's reward, on the y-axis. Each plot shows results in different environments, with different adversarial attacks on the x-axis. We show both the raw adversary score, as well as the adversary score adjusted for detection rates of different adversarial attacks (see Figure 5). While the SA-MDP and MNP benchmark attacks achieve higher unadjusted scores, their high detection rates result in significantly lower adjusted scores.

Relative adversary scores. We now investigate the adversaries' relative performance using the scalar *adversary score*, i.e. the the resultant (normalized) reduction in the victim's return. We find that the adversary score is generally higher for SA-MDP attacks than for R-illusory attacks and *perfect illusory attacks* (see Figure 4). This is expected, as SA-MDP attacks purely minimise the victim reward, while R-illusory attacks trade off between minimizing the reward of the victim and remaining undetected. Further, perfect illusory attacks achieve a lower adversary score than R-illusory attacks. This is likewise to be expected, as perfect illusory attacks do not explicitly lower the victim reward.

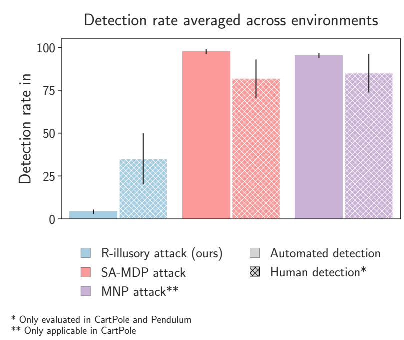

Figure 5: Different adversarial attacks are shown on the x-axis, with detection rates on the y-axis. We see that both the automated detector as well as human subjects are able to detect SA-MDP and MNP attacks, while R-illusory attacks are less likely to be detected. Perfect illusory attacks are excluded here as they are undetectable.

Detection-adjusted adversary scores. We now adjust adversary scores to include the outcome of the automated detector Δ by setting the adversary's return to zero across episodes classified as attacked. This reflects a middle ground across different scenarios in which adversary detection could trigger victim contingency options ranging from no action, to test-time termination, major security escalation, and adversary persecution. Specifically, we adjust the adversary score by multiplying it by $1 - \mathbb{P}(\text{detection})$. Figure 4 shows that R-illusory attacks result in the highest average adversary score. In contrast, detection-adjusted adversary scores for state-of-the-art attacks are close to zero as expected from their high empirical detectability (see Fig. 5).

5. Conclusion and future work

This paper introduces a novel class of observation-space adversarial attacks, *illusory attacks*, which aim to minimize statistical detectability. We study both automated detection and detection by humans of existing observation-space adversarial attacks and illusory attacks.

We expect the potential positive impact of enabling adversarial defense systems to counter illusory attacks to outweigh the potential negative consequences associated with studying enhanced adversarial attacks. However, it should be acknowledged we assume the availability of contingency options for victim agents, which may not always hold true in real-world scenarios. Moreover, our experimental investigations are confined to simulated environments, necessitating further exploration in more intricate real-world domains.

Future research should conduct comprehensive theoretical analysis of the Nash equilibria within the two-player zero-sum game introduced by the illusory attack framework. Furthermore, efforts are required to develop more effective detection mechanisms and robustification techniques that are applicable to real-world environments. An equally significant aspect of detection is gaining a deeper understanding of the human capability to perceive and identify (illusory) adversarial attacks. We ultimately aim to demonstrate the viability of illusory attacks and the corresponding defense strategies in real-world settings, particularly in mixed-autonomy scenarios.

References

[1] Abdelfattah, M., Yuan, K., Wang, Z. J., and Ward, R. Towards Universal Physical Attacks On Cascaded Camera-Lidar 3d Object Detection Models. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3592–3596, September 2021. doi: 10.1109/ICIP42928.2021.9506016. ISSN: 2381-8549.

[2] 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.

[3] Bakhtin, A., Wu, D. J., Lerer, A., Gray, J., Jacob, A. P., Farina, G., Miller, A. H., and Brown, N. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. *arXiv preprint arXiv:2210.05492*, 2022.

[4] Barto, A., Lewis, R. L., and Singh, S. Where Do Rewards Come From. 2009.

[5] Basseville, M., Nikiforov, I. V., et al. *Detection of abrupt changes: theory and application*, volume 104. prentice Hall Englewood Cliffs, 1993.

[6] Bellman, R. Dynamic programming and stochastic control processes. *Information and Control*, 1(3):228–239, September 1958. ISSN 0019-9958. doi: 10.1016/S0019-9958(58)80003-0.

[7] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., and Zieba, K. End to End Learning for Self-Driving Cars, April 2016. URL http://arxiv.org/abs/1604.07316. arXiv:1604.07316 [cs].

[8] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.

[9] Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q. A., Fu, K., and Mao, Z. M. Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pp. 2267–2281, New York, NY, USA, November 2019. Association for Computing Machinery. ISBN 978-1-4503-6747-9. doi: 10.1145/3319535.3339815. URL https://dl.acm.org/doi/10.1145/3319535.3339815.

[10] Cao, Y., Wang, N., Xiao, C., Yang, D., Fang, J., Yang, R., Chen, Q. A., Liu, M., and Li, B. Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks. pp. 176–194. IEEE Computer Society, May 2021. ISBN 978-1-72818-934-5. doi: 10.1109/SP40001.2021.00076. URL https://www.computer.org/csdl/proceedings-article/sp/2021/893400b302/1t0x9btzenu.

[11] Cazorla, L., Alcaraz, C., and Lopez, J. Cyber Stealth Attacks in Critical Information Infrastructures. *IEEE Systems Journal*, 12(2):1778–1792, June 2018. ISSN 1937-9234. doi: 10.1109/JSYST.2015.2487684. Conference Name: IEEE Systems Journal.

[12] Chen, S.-T., Cornelius, C., Martin, J., and Chau, D. H. P. ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector. In Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., and Ifrim, G. (eds.), *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pp. 52–68, Cham, 2019. Springer International Publishing. ISBN 978-3-030-10925-7. doi: 10.1007/978-3-030-10925-7_4.

[13] 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.

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

[15] Das, N., Shanbhogue, M., Chen, S.-T., Hohman, F., Li, S., Chen, L., Kounavis, M. E., and Chau, D. H. SHIELD: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 196–204, New York, NY, USA, July 2018. Association for Computing Machinery. ISBN 978-1-4503-5552-0. doi: 10.1145/3219819.3219910. URL https://dl.acm.org/doi/10.1145/3219819.3219910.

[16] de Witt, C. S. and Hornigold, T. Stratospheric Aerosol Injection as a Deep Reinforcement Learning Problem, May 2019. URL http://arxiv.org/abs/1905.07366. arXiv:1905.07366 [physics, stat].

[17] Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

[18] Dennis, M., Jaques, N., Vinitsky, E., Bayen, A., Russell, S., Critch, A., and Levine, S. Emergent complexity and zero-shot transfer via unsupervised environment design. *Advances in neural information processing systems*, 33:13049–13061, 2020.

[19] Doob, J. L. Stochastic Processes and Statistics. *Proceedings of the National Academy of Sciences of the United States of America*, 20(6): 376–379, June 1934. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1076423/.

[20] Duan, R., Ma, X., Wang, Y., Bailey, J., Qin, A. K., and Yang, Y. Adversarial Camouflage: Hiding Physical-World Attacks With Natural Styles. pp. 997–1005. IEEE Computer Society, June 2020. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00108. URL https://www.computer.org/csdl/proceedings-article/cvpr/2020/716800a997/1m3oiKjRrX2.

[21] 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.

[22] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust Physical-World Attacks on Deep Learning Visual Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, June 2018. doi: 10.1109/CVPR.2018.00175. ISSN: 2575-7075.

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

[24] Ha, D. and Schmidhuber, J. World Models. March 2018. doi: 10.5281/zenodo.1207631.

[25] 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.

[26] Halpern, J. Y. and Moses, Y. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, July 1990. ISSN 0004-5411. doi: 10.1145/79147.79161. URL https://dl.acm.org/doi/10.1145/79147.79161.

[27] He, K., Gkioxari, G., Dollar, P., and Girshick, R. Mask R-CNN. pp. 2961–2969, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html.

[28] Hu, Y.-C.-T., Chen, J.-C., Kung, B.-H., Hua, K.-L., and Tan, D. S. Naturalistic Physical Adversarial Patch for Object Detectors. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7828–7837, October 2021. doi: 10.1109/ICCV48922.2021.00775. ISSN: 2380-7504.

[29] Huang, L., Gao, C., Zhou, Y., Xie, C., Yuille, A. L., Zou, C., and Liu, N. Universal Physical Camouflage Attacks on Object Detectors. pp. 717–726. IEEE Computer Society, June 2020. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00080. URL https://www.computer.org/csdl/proceedings-article/cvpr/2020/716800a717/1m3okhiXfzy.

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

[31] Huang, Y. and Zhu, Q. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *International Conference on Decision and Game Theory for Security*, pp. 217–237. Springer, 2019.

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

[33] Hutter, M. *Universal Artificial Intellegence*. Texts in Theoretical Computer Science An EATCS Series. Springer, Berlin, Heidelberg, 2005. ISBN 978-3-540-22139-5 978-3-540-26877-2. doi: 10.1007/b138233. URL http://link.springer.com/10.1007/b138233.

[34] Ilahi, I., Usama, M., Qadir, J., Janjua, M. U., Al-Fuqaha, A., Hoang, D. T., and Niyato, D. Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *IEEE Transactions on Artificial Intelligence*, 3(2):90–109, 2021.

[35] Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, May 1998. ISSN 0004-3702. doi: 10.1016/S0004-3702(98)00023-X.

[36] Khintchine, A. Korrelationstheorie der stationären stochastischen Prozesse. *Mathematische Annalen*, 109 (1):604–615, December 1934. ISSN 1432-1807. doi: 10.1007/BF01449156. URL https://doi.org/10.1007/BF01449156.

[37] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[39] Kovařík, V., Schmid, M., Burch, N., Bowling, M., and Lisý, V. Rethinking formal models of partially observable multiagent decision making. *Artificial Intelligence*, 303:103645, February 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103645. URL https:

//www.sciencedirect.com/science/article/pii/S000437022100196X.

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

[41] Lai, T. L. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information theory*, 44(7):2917–2929, 1998.

[42] Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., and Beijbom, O. PointPillars: Fast Encoders for Object Detection From Point Clouds. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12689–12697, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.01298. URL https://ieeexplore.ieee.org/document/8954311/.

[43] Langner, R. Stuxnet: Dissecting a Cyberwarfare Weapon. *IEEE Security & Privacy*, 9(3):49–51, May 2011. ISSN 1558-4046. doi: 10.1109/MSP.2011.67. Conference Name: IEEE Security & Privacy.

[44] Lanier, J. B., McAleer, S., Baldi, P., and Fox, R. Feasible adversarial robust reinforcement learning for underspecified environments. *arXiv preprint arXiv:2207.09597*, 2022.

[45] Leike, J. *Nonparametric General Reinforcement Learning*. PhD thesis, Australian National University, November 2016. URL http://arxiv.org/abs/1611.08944. arXiv:1611.08944 [cs].

[46] Li, S., Neupane, A., Paul, S., Song, C., Krishnamurthy, S. V., Chowdhury, A. K. R., and Swami, A. Adversarial Perturbations Against Real-Time Video Classification Systems. In *Proceedings 2019 Network and Distributed System Security Symposium*, 2019.

[47] Liang, Y., Sun, Y., Zheng, R., and Huang, F. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

[48] Lin, Y.-C., Liu, M.-Y., Sun, M., and Huang, J.-B. Detecting adversarial attacks on neural network policies with visual foresight. *arXiv preprint arXiv:1710.00814*, 2017.

[49] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. May 2023. URL https://openreview.net/forum?id=rJzIBfZAb.

[50] McAleer, S. M., Farina, G., Lanctot, M., and Sandholm, T. ESCHER: Eschewing Importance Sampling in Games by Computing a History Value Function to Estimate Regret. February 2023. URL https://openreview.net/forum?id=35QyoZv8cKO.

[51] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, July 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3059968. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[52] 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.

[53] Monahan, G. E. A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms. *Management Science*, 28(1):1–16, 1982. ISSN 0025-1909. URL https://www.jstor.org/stable/2631070. Publisher: INFORMS.

[54] Nash, J. F. Equilibrium Points in N-Person Games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1):48–49, January 1950. ISSN 0027-8424. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1063129/.

[55] Neven, D., De Brabandere, B., Georgoulis, S., Proesmans, M., and Van Gool, L. Towards End-to-End Lane Detection: an Instance Segmentation Approach, February 2018. URL http://arxiv.org/abs/1802.05591. arXiv:1802.05591 [cs].

[56] Oikarinen, T., Zhang, W., Megretski, A., Daniel, L., and Weng, T.-W. Robust deep reinforcement learning through adversarial loss. *Advances in Neural Information Processing Systems*, 34, 2021.

[57] Page, E. S. Continuous Inspection Schemes. *Biometrika*, 41(1/2):100–115, 1954. ISSN 0006-3444. doi: 10.2307/2333009. URL https://www.jstor.org/stable/2333009. Publisher: [Oxford University Press, Biometrika Trust].

[58] 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.

[59] Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

[60] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. pp. 779–788. IEEE Computer Society, June

2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.91. ISSN: 1063-6919.

[61] Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html.

[62] Ross, S. M. *Introduction to Probability Models*. Academic Press, Amsterdam ; Boston, 10th edition edition, December 2009. ISBN 978-0-12-375686-2.

[63] Russo, A. and Proutiere, A. Balancing detectability and performance of attacks on the control channel of markov decision processes. In *2022 American Control Conference (ACC)*, pp. 2843–2850. IEEE, 2022.

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

[65] Samangouei, P., Kabkab, M., and Chellappa, R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. May 2023. URL https://openreview.net/forum?id=BkJ3ibb0-.

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

[67] Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. Does the Markov Decision Process Fit the Data: Testing for the Markov Property in Sequential Decision Making, February 2020. arXiv:2002.01751 [cs, stat].

[68] Shi, S., Wang, X., and Li, H. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–779, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00086. URL https://ieeexplore.ieee.org/document/8954080/.

[69] Sokota, S., D'Orazio, R., Kolter, J. Z., Loizou, N., Lanctot, M., Mitliagkas, I., Brown, N., and Kroer, C. A Unified Approach to Reinforcement Learning, Quantal Response Equilibria, and Two-Player Zero-Sum Games, April 2023. URL http://arxiv.org/abs/2206.05825. arXiv:2206.05825 [cs].

[70] Sun, J., Cao, Y., Chen, Q. A., and Mao, Z. M. Towards robust LiDAR-based perception in autonomous driving: general black-box adversarial sensor attack and countermeasures. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, pp. 877–894, USA, August 2020. USENIX Association. ISBN 978-1-939133-17-5.

[71] Sun, J., Zhang, T., Xie, X., Ma, L., Zheng, Y., Chen, K., and Liu, Y. Stealthy and Efficient Adversarial Attacks against Deep Reinforcement Learning, May 2020. arXiv:2005.07099 [cs].

[72] Sun, Y., Zheng, R., Liang, Y., and Huang, F. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep rl. In *International Conference on Learning Representations*, 2021.

[73] Sutton, R. S. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, July 1991. ISSN 0163-5719. doi: 10.1145/122344.122377. URL https://dl.acm.org/doi/10.1145/122344.122377.

[74] Sutton, R. S. The Quest for a Common Model of the Intelligent Decision Maker, June 2022.

[75] 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.

[76] Szepesvári, C. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing, Cham, 2010. ISBN 978-3-031-00423-0 978-3-031-01551-9. doi: 10.1007/978-3-031-01551-9. URL https://link.springer.com/10.1007/978-3-031-01551-9.

[77] Tartakovsky, A., Nikiforov, I., and Basseville, M. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

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

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

[80] Tu, J., Ren, M., Manivasagam, S., Liang, M., Yang, B., Du, R., Cheng, F., and Urtasun, R. Physically Realizable Adversarial Examples for LiDAR Object Detection. pp. 13713–13722. IEEE Computer Society, June 2020. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.01373. URL https://www.computer.org/csdl/proceedings-article/cvpr/2020/716800n3713/1m3o75on8VG.

[81] Von Neumann, J. and Morgenstern, O. *Theory of games and economic behavior*. Theory of games and

economic behavior. Princeton University Press, Princeton, NJ, US, 1944. Pages: xviii, 625.

[82] Vyas, S., Hannay, J., Bolton, A., and Burnap, P. P. Automated Cyber Defence: A Review, March 2023. URL http://arxiv.org/abs/2303.04926. arXiv:2303.04926 [cs].

[83] Wang, Y., Chao, W.-L., Garg, D., Hariharan, B., Campbell, M., and Weinberger, K. Q. Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. pp. 8437–8445. IEEE Computer Society, June 2019. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00864. URL https://www.computer.org/csdl/proceedings-article/cvpr/2019/329300i437/1gyrjqe4Oac.

[84] Werbos, P. J. Applications of advances in nonlinear sensitivity analysis. In *System Modeling and Optimization: Proceedings of the 10th IFIP Conference New York City, USA, August 31–September 4, 1981*. Springer, 2005.

[85] 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.

[86] Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. Mitigating Adversarial Effects Through Randomization. May 2023. URL https://openreview.net/forum?id=Sk9yuql0Z.

[87] Xu, W., Evans, D., and Qi, Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Proceedings 2018 Network and Distributed System Security Symposium*, 2018. doi: 10.14722/ndss.2018.23198. URL http://arxiv.org/abs/1704.01155. arXiv:1704.01155 [cs].

[88] Yang, B., Luo, W., and Urtasun, R. PIXOR: Real-time 3D Object Detection from Point Clouds. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7652–7660, Salt Lake City, UT, USA, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00798. URL https://ieeexplore.ieee.org/document/8578896/.

[89] Yu, H. and Bertsekas, D. P. On near Optimality of the Set of Finite-State Controllers for Average Cost POMDP. *Mathematics of Operations Research*, 33 (1):1–11, 2008. ISSN 0364-765X. URL https://www.jstor.org/stable/25151838. Publisher: INFORMS.

[90] Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020.

[91] Zhang, H., Chen, H., Boning, D., and Hsieh, C.-J. Robust Reinforcement Learning on State Observations with Learned Optimal Adversary, January 2021.

[92] Zhou, Y. and Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, Salt Lake City, UT, USA, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00472. URL https://ieeexplore.ieee.org/document/8578570/.

[93] Zhu, Y., Miao, C., Zheng, T., Hajiaghajani, F., Su, L., and Qiao, C. Can We Use Arbitrary Objects to Attack LiDAR Perception in Autonomous Driving? In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pp. 1945–1960, New York, NY, USA, November 2021. Association for Computing Machinery. ISBN 978-1-4503-8454-4. doi: 10.1145/3460120.3485377. URL https://dl.acm.org/doi/10.1145/3460120.3485377.

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

A. Appendix

B. Related work

The **adversarial attack** literature originates in image classification [75], where attacks commonly need to be visually imperceptible. Visual imperceptibility is commonly proxied by simple pixel-space minimum-norm perturbation (MNP) constraints [23, 49]. Several defenses against MNP attacks have been proposed [15, 87, 65, 86].

MNP attacks have been extended to **adversarial attacks on sequential decision-making agents** [13, 34, 58]. In the sequential MNP framework, the adversary can modify the victim's observations up to a step- or episode-wise perturbation budget, both in white-box, as well as in black-box settings. Zhang et al. [90] and Sun et al. [72] use reinforcement learning to learn adversarial policies that require only black-box access to the victim policy. Assuming a different black-box setting, Hussenot et al. [32] introduce a class of adversaries for which a unique mask is precomputed and added to the agent observation at every time step. Our framework differs from these previous works in that it takes into account the temporal consistency of observation sequences.

Work towards **robust sequential-decision making** uses techniques such as randomized smoothing [40, 85], test-time hardening by computing confidence bounds [21], training with adversarial loss functions [56], and co-training with adversarial agents [91, 18, 44]. We compare against and build upon this work.

Various strands of research in cyber security concern **adversarial patch (AP) attacks** that do not require access to all the sensor pixels, and commonly assume that the attack target can be physically modified [22, 10]. AP attack targets include cameras [22, 12, 20, 29, 28], LiDAR [70, 9, 93, 80], and multi-sensor fusion mechanisms [10, 1]. Our *illusory* attack framework differs from both MNP and AP attacks in that it is not restricted to patches, does not require perturbation budgets, and introduces explicit detectability constraints. In contrast, our work explores both defenses based on learnt world models, and human detectors rather than hand-crafted detectors that require domain knowledge.

Another body of work focuses on **detection and detectability of learnt adversarial attacks on sequential decision makers**. Lin et al. [48] develop an action-conditioned frame module that allows agents to detect adversarial attacks by comparing both the module's action distribution with the realised action distribution. Tekgul et al. [78] detect adversaries by evaluating the feasibility of past action sequences. Li et al. [46], Sun et al. [71], Huang & Zhu [31] focus on the detectability of adversarial attacks but without considering notions of stochastic equivalence between observation processes. Perhaps most closely related to our work, Russo & Proutiere [63] study action-space attacks on low-dimensional stochastic control systems and consider information-theoretic detection [5, 41, 77] based on stochastic equivalence between the resulting trajectories. We instead investigate high-dimensional observation-space attacks, and consider learned detectors, as well as humans.

B.1. Proof of Theorem 3.2

We first restate Theorem 3.2 in a slightly more precise way. Consider a POMDP $\mathcal{E}_e := \langle \mathcal{S}', \mathcal{A}, \Omega, \mathcal{O}', p', r, \gamma \rangle$ with finite horizon T, a state space $\mathcal{S}' := (\mathcal{S} \times \mathcal{A} \times \Omega)^T$, deterministic observation function $\mathcal{O}' : \mathcal{S}' \mapsto \Omega$, and stochastic state transition function $p' : \mathcal{S}' \times \mathcal{A} \mapsto \mathcal{P}(\mathcal{S}')$. Then, for any $\pi_v : \mathcal{H}^*_{\setminus r} \mapsto \mathcal{P}(\mathcal{A})$ and $\nu : \mathcal{S} \times \mathcal{H}^*_{\setminus r} \mapsto \mathcal{P}(\Omega)$, we can define corresponding p' and $\mathcal{O}'$ such that the reward and observation processes cannot be distinguished by the victim.

Recall that the semantics of $\mathcal{E}^\pi_\nu$ are as follows: Fix a victim policy $\pi : \mathcal{H}^*_{\setminus r} \mapsto \mathcal{P}$ from the space of all possible sampling policies Π. At time $t = 0$, we sample an initial state $s_0 \sim p(\cdot|\emptyset)$. The adversary then samples an observation $o_0 \sim \nu(\cdot|s_0)$ which is emitted to the victim. The victim takes an action $a_0 \sim \pi(\cdot|o_0)$, upon which the state transitions to $s_1 \sim p(\cdot|s_0, a_0)$ and the victim receives a reward $r_1 \sim (\cdot|s_0, a_0)$. At time $t > 0$, the victim has accumulated a history $h_t := o_0 a_0 r_1 \dots o_t$, on which $o_t \sim \nu(\cdot|s_t, h_{t\setminus r})$ conditions.

Proof. Now consider an equivalent POMDP formulation. Define p' as the following sequential stochastic process: At time $t = 0$, first sample $s_0 \sim p(\cdot|\emptyset)$. Then sample $o_0 \sim \nu(\cdot|s_0)$, and define $s'_0 := p'(\emptyset) := (s_0, o_0)$. For any $t > 0$, first sample $s_t \sim p(\cdot|s_{t-1}, a_{t-1})$, then $o_t \sim \nu(\cdot|s_{\leq t}, a_{<t}, o_{<t})$ and define $s'_t := p'(s'_{t-1}, s_t, o_t, a_{t-1})$. We finally define $\mathcal{O}(s'_t) := proj_o(s'_t) := o_t$, where we indicate that o_t is stored in s'_t by using an explicit projection operator $proj_o$. Clearly, under any sampling policy π, the observation and reward processes induced by $\mathcal{E}_e$ and $\mathcal{E}^{\pi_v}_\nu$ are identical as $T \to \infty$. This renders the reward and observation processes identical in both environments. Note that, as $T \to \infty$, $\mathcal{E}_e$'s state space grows infinitely large. □

B.2. Proof on Theorem 3.4

Given an MDP $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$, we consider instead its canonical embedding (see Section 2) into the family of POMDPs $\mathcal{E} :=$ pomdp$(\mathcal{M})$ with state space $\mathcal{S}$ and observation function $\mathcal{O} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$.

Given sampling access to another POMDP $\mathcal{K}$, we wish to prove that the protocol defined in Theorem 3.4 establishes stochastic equivalence between the observation processes of $\mathcal{E}$ and $\mathcal{K}$ (see 3.3) under an arbitrary sampling policy $\pi : \mathcal{H}^* \mapsto \mathcal{P}(\mathcal{A})$.

By definition, $\mathcal{M}$'s state-transition function obeys the Markov property. As $\mathcal{O}$ is the identity function, $\mathcal{E}^\pi$'s observation process hence also obeys the Markov property. By definition of stochastic equivalence (see 3.3), for $\mathcal{E}^\pi$'s observation process to be stochastically equivalent to $\mathcal{K}^\pi$'s, the observation processes of $\mathcal{K}^\pi$ and $\mathcal{E}^\pi$ have to fulfill Definition 3.3. As $\mathcal{E}^\pi$'s observation process has the Markov property, $\mathcal{K}^\pi$ hence also needs to have the Markov property (Step 3 in Theorem 3.4). If, in addition, $\mathcal{K}^\pi$ fulfills Steps 1 and 2, then this implies stochastic equivalence between $\mathcal{K}^\pi$ and $\mathcal{E}^\pi$.

B.3. Proof on the existence of perfect illusory attacks

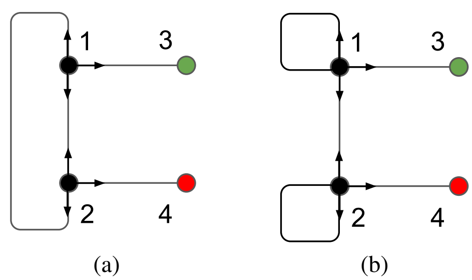

Figure 6: An environment for which perfect illusory attacks do exist (left), and one for which they do not exist (right).

Proof that perfect illusory attacks exist. We consider an example MDP (see Figure 6a) where a victim starts in node 1 or 2 each with probability $\frac{1}{2}$ and can go *up*, *down*, or *right* in both states 1 and 2. The episode terminates immediately with a return of 0 should the victim reach state 4. Otherwise, the victim receives a reward of $+1$ if it reaches state 3 within a maximum of 2 steps. The optimal victim policy is therefore to take paths $1 \rightarrow 3$ if starting in state 1, and take one of the two possible paths $2 \rightarrow 1 \rightarrow 3$ otherwise. The victim observes the labelled state graph, as well as its current state label. Clearly, choosing $\nu(1) = 2$ and $\nu(2) = 1$ constitutes a perfect illusory attack in this environment. $\square$

Proof that perfect illusory attacks don't always exist. To show that some environments do not admit perfect illusory attacks, consider the modified environment in Figure 6b. Here, clearly a timestep-conditioned victim policy that takes the action sequence $\langle \text{up}, \text{right} \rangle$ independently of observations cannot be perfectly attacked. $\square$

B.4. Two stochastic processes can have the same states and state-transition function, but not be equal

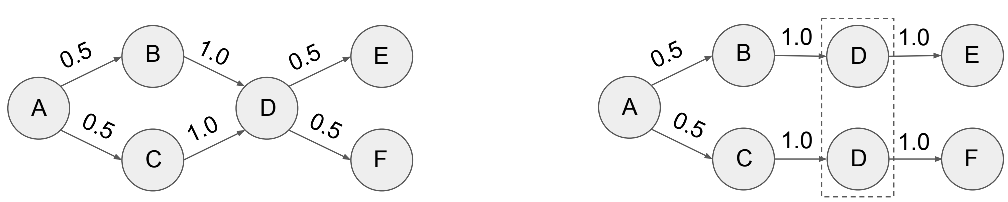

Figure 7: Left: An unattacked Markov Decision Process with 6 states. Right: A decision process with the same state-transition function, but long-term correlations.

Two stochastic processes can have the same states and state-transition functions, but still not be equal. Figure 7 illustrates this with a simple example: The attacked process on the right transitions to state E whenever it has been in B prior, and transitions to F whenever it has been in C prior. This process has the exact same state-transition function as the process on the left, however, the processes are not equivalent.

B.5. Implementation overview and additional discussion.

For all four evaluation environments, we succeed in implementing *perfect illusory attacks* (see Definition 3.5) by first constructing an attacked initial state distribution $p(\cdot|\emptyset)$ that exploits environment-specific symmetries. We then sample the initial attacked observations o_0 from the attacked initial state distribution and generate subsequent transitions using the unattacked state transition function $p(\cdot|o_{t-1}, a_{t-1})$ where a_{t-1} is the action taken at the last time step. For the Algorithm used and more details see also Appendix B.9. In contrast to perfect illusory attacks, *R-illusory attacks* are learned end-to-end using reinforcement learning. As detailed in Algorithm 1, the adversary's reward is given by the negative victim reward plus an *illusory reward* that incentivises attacked observations to be aligned with the unattacked state-transition function p. We choose the illusory reward to be proportional to the L_∞-norm of the distance between the next state according to p and the attacked observation, acting as a proxy for the distance between the unattacked and attacked distributions. A hyperparameter λ trades off between victim reward and illusory reward. We ran a grid search over λ and found that for $1 < \lambda < 1000$ results are mostly insensitive to λ, while $\lambda < 1$ and $\lambda > 1000$ result in either non-illusory attacks, or unattacked observations, respectively.

State-of-the-art adversarial attacks. We consider MNP [40] and SA-MDP [91] adversarial attacks together to be sufficiently representative of the state-of-the-art (see Section 2). Following recent works, we evaluate both MNP and SA-MDP attacks for perturbation budgets $\beta \in \{0.05, 0.2\}$, where β is defined relative to normalised observation vectors. For a fair comparison, we impose the same βs on R-illusory attacks.

Robustification and Detectors. We test the defenses *randomised smoothing (RS)* [14, 40] and *adversarial pre-training (ATLA)* [91]. In RS, Gaussian noise is i.i.d. sampled at each time step and added to the victim's observation o_t. In adversarial pre-training, the victim policy is learned through co-training with the adversary. In ATLA, the victim is co-trained with an SA-MDP adversary. We also consider an ablation of ATLA where we pre-train the victim agent with an R-illusory adversary (which we refer to as *ATLA abl.*). For each of the four environments, we train a small neural network-based *world model* [84, 24] $\hat{p}(\cdot|s, a)$ using trajectory rollouts from the unattacked environment. The world model approximates the environment's state-transition function and can be used to approximate the detection rule used for quickest change detection (see Appendix B.8). The detector Δ uses $\hat{p}$ to check whether, for a given observed transition (o_t, a_t, o_{t+1}) it holds that the distance between the next predicted observation and the next observation is smaller than a threshold c, i.e. whether $\|\hat{p}(o_t, a_t) - o_{t+1}\|_\infty \leq c$. If this does not hold, the trajectory is classified as attacked. We evaluate Δ on a held-out set of train-time trajectories and tune the detection threshold c to achieve a false positive (FP) rate of 3%.

B.6. Detection by human inspection

We investigate whether humans unfamiliar with adversarial attacks can detect adversarial attacks in simple and easy-to-understand environments. We found *CartPole* and *Pendulum*, in contrast to Hopper and HalfCheetah, to be immediately accessible to participants without the need for additional training. To this end, we perform a controlled study with $n = 10$ human participants. Participants were first shown an unattacked *introduction video* for both *CartPole* and *Pendulum*, exposing them to environment-specific dynamics. Participants were then shown a random set of videos containing both videos of unattacked and attacked trajectories. For each video, participants were asked to indicate whether they believed that the video was unsuspicious, with the prompt *"the system shown in the video was [not] the same as the one from the introduction video"*. This phrasing was chosen so that participants would not be primed on the concept of illusory attacks (see details in Appendix B.11).

We omit perfect illusory attacks in the study as these, by definition, cannot be detected. We found that participants classified MNP and SA-MDP attacks as suspicious with high accuracy (see *Human detection* in Figure 5). In contrast, participants were almost equally likely to classify videos of unattacked and R-illusory attacked trajectories as unsuspicious. In fact, at a confidence level of 95%, the hypothesis *participants are equally likely to classify an unattacked sequence as attacked as to classify an R-illusory attacked sequence as attacked* cannot be rejected. Our findings suggest that humans are unable to detect R-illusory attacks from short observation sequences in our simple environments. See Appendix B.11 for full results and the corresponding z-test statistic.

Robustification and Detectors. We test the defenses *randomised smoothing (RS)* [14, 40] and *adversarial pre-training (ATLA)* [91]. In RS, Gaussian noise is i.i.d. sampled at each time step and added to the victim's observation o_t. In adversarial pre-training, the victim policy is learned through co-training with the adversary. In ATLA, the victim is co-trained with

an SA-MDP adversary. We also consider an ablation of ATLA where we pre-train the victim agent with an R-illusory adversary (which we refer to as *ATLA abl.*). For each of the four environments, we train a small neural network-based *world model* [84, 24] $\hat{p}(\cdot|s,a)$ using trajectory rollouts from the unattacked environment. The world model approximates the environment's state-transition function and can be used to approximate the detection rule used for quickest change detection (see Appendix B.8). The detector Δ uses $\hat{p}$ to check whether, for a given observed transition (o_t, a_t, o_{t+1}) it holds that the distance between the next predicted observation and the next observation is smaller than a threshold c, i.e. whether $\|\hat{p}(o_t, a_t) - o_{t+1}\|_\infty \leq c$. If this does not hold, the trajectory is classified as attacked. We evaluate Δ on a held-out set of train-time trajectories and tune the detection threshold c to achieve a false positive (FP) rate of 3%.

B.7. R-illusory attacks adversarial training

Algorithm 1 R-illusory attacks adversarial training

Input: environment *env* with transition function p, illusory reward weight λ, victim policy π_v, number of training episodes N.
Init. adversary policy ν_ψ with parameters ψ.
while episode $<$ N **do**
  $t = 0$
  $s_0 =$ env.reset()
  $o_0 = \nu_\psi(s_0)$
  $a_0 = \pi_v(o_0)$
  $o_1, r_1,$ *done* $=$ env.step(a_0)
  $r_1^{\text{adv}} = -r_1$
  done $=$ *False*
  while not *done* **do**
    $t = t + 1$
    $o_t = \nu_\psi(s_t)$
    $a_t = \pi_v(o_t)$
    $s_{t+1}, r_{t+1},$ *done* $=$ env.step(a_t)
    $r_{t+1}^{\text{adv}} = -r_{t+1} - \lambda \cdot \|o_t - p(o_{t-1}, a_{t-1})\|_\infty$
  end while
  Update ν_ψ from tuples $(s_t, o_t, r_{t+1}^{\text{adv}}, s_{t+1})$.
end while

B.8. Detector used in experiments

We assume that the victim is trained in the unattacked environment $\mathcal{E}$ for k episodes, each consisting of n steps. During training, the agent records observed environment transition tuples denoted as $t_i = (s_i, a_i, s_{i+1})$, which are stored in a set $\mathcal{D}_\mathcal{E} = t_{i=0}^{i=k*(n-1)}$. These unattacked transitions in $\mathcal{D}_\mathcal{E}$ are used to learn an approximation of the state-transition function denoted as $\hat{p}_\mathcal{E}(s_t|s_{t-1}, a_{t-1})$. To implement $\hat{p}_\mathcal{E}$, we employ a Multi-Layer Perceptron with two hidden layers of size 10. The model is trained using an $l2$ loss for 10 epochs, with a learning rate of 0.001 and the ADAM optimizer [37].

At each time step t, the test statistic z_t, utilized in CUSUM change point detection methods [5, 41, 77], is computed based on the transition tuple consisting of the last attacked observation o_{t-1}, the last action taken a_{t-1}, and the current attacked observation o_t. In other words, the transition tuple is represented as (o_{t-1}, a_{t-1}, o_t). To compute z_t, we require the train-time transition distribution $p_\mathcal{E}$ and the test-time transition distribution $p_{\mathcal{E}'}$, which are both unknwon and must be estimated.

The test statistic z_t is defined as the logarithm of the ratio between the probability of the transition tuple under the test-time distribution and the train-time distribution:

$$z_t(o_{t-1}, a_{t-1}, o_t) = \ln \frac{p_{\mathcal{E}'}(o_t \mid a_{t-1}, o_{t-1})}{p_\mathcal{E}(o_t \mid a_{t-1}, o_{t-1})}. \tag{4}$$

Given that state transitions in the provided environments are deterministic, the estimated probabilities are either 0 or δ. Moreover, the estimated transition probabilities of observed transition tuples for the test-time process are δ. Therefore, it follows that $p_{\mathcal{E}}(o_t \mid a_{t-1}, o_{t-1}) \in \{0, \delta\}\ \forall\ t$ and $p_{\mathcal{E}'}(o_t \mid a_{t-1}, o_{t-1}) = \delta\ \forall\ t$. Consequently, $z_t(o_{t-1}, a_{t-1}, o_t) \in \{0, \infty\}\ \forall\ t$.

CUSUM-based methods rely on accumulating the test statistic z_t over multiple time steps and classify a sequence as attacked if the sum exceeds a threshold. In our case, since the test statistic is either 0 or ∞, the CUSUM testing procedure simplifies to classifying a sequence as attacked when $z_t = \infty$. Furthermore, $z_t = \infty$ if $p_{\mathcal{E}}(o_t \mid a_{t-1}, o_{t-1}) = 0$.

To determine $p_{\mathcal{E}}(o_t \mid a_{t-1}, o_{t-1})$, we utilize the learned state-transition model. If the l_2 distance between the predicted next observation $\hat{p}_{\mathcal{E}}(o_{t-1}, a_{t-1})$ and and the given next observation o_t exceeds a threshold c, we set $p_{\mathcal{E}}(o_t \mid a_{t-1}, o_{t-1})$ to 0. The threshold c accounts for the inaccuracy of $\hat{p}_{\mathcal{E}}$. To determine an appropriate value for c, we evaluate $\hat{p}_{\mathcal{E}}$ on a held-out set of training observations and tune c to achieve a false positive rate of 3%. We assess the accuracy of detecting adversarial attacks across all scenarios presented in Table B.10.1.

B.9. Perfect illusory attacks implementation

We implement perfect illusory attacks in *CartPole* and *Pendulum* as detailed in Algorithm 2. The first observation o_0 is set to the negative of the true first state sampled from the environment, i.e. $o_0 = -o_0$. Note that in *HalfCheetah* and *Hopper* the initial state distribution is not centered around the origin, we hence first subtract the offset, and then compute the negative of the observation and add the offset again. As the distribution over initial states is symmetric in all environments (after removing the offset), this approach satisfies the conditions of a perfect illusory attack (see Definition 3.5). We provide videos of the generated perfect illusory attacks in the supplementary material in the respective folder and show an illustration of a perfect illusory attack in Figure 2.

Algorithm 2 Perfect illusory adversarial training

Input: environment env, environment transition function t whose initial state distribution $p(\cdot|\emptyset)$ is symmetric with respect to the point $p_{symmetry}$ in $\mathcal{S}$, victim policy π_v.
$k = 0$
$s_0 = env$.reset()
$o_0 = -(s_0 - p_{symmetry}) + p_{symmetry}$
$a_0 = \pi_v(o_0)$
_, $done = env$.step(a_0)
while not *done* **do**
  $k = k + 1$
  $o_k \sim t(o_{k-1}, a_{k-1})$
  $a_k = \pi_v(o_k)$
  _, $done = env$.step(a_k)
end while

B.10. Learning R-illusory attacks with reinforcement learning

We next describe the algorithm used to learn R-illusory attacks and the training procedures used to compute the results in Table B.10.1. We use the *CartPole*, *Pendulum*, *HalfCheetah* and *Hopper* environments as given in Brockman et al. [8]. We shortened the episodes in *Hopper* and *HalfCheetah* to 300 steps to speed up training. The transition function is implemented using the physics engines given in all environments. We normalise observations by the maximum absolute observation. We train the victim with PPO [66] and use the implementation of PPO given in Raffin et al. [59], while not making any changes to the given hyperparameters. In both environments we train the victim for 1 million environment steps. We implement the ATLA [91] victim by co-training it with an adversary agent, and follow the original implementation of the authors [5]. We implement the ablation of ATLA [91] that trains the victim with an illusory adversary by replacing the SA-MDP adversary with an R-illusory attack adversary, which is implemented as stated in algorithm 1. For co-training, we alternate between training the victim and the adversary agent every 400 environment steps. This parameter was chosen in a small evaluation study as it yields non-oscillating behaviour. We further investigated different ratios between training steps of the adversary

[5] https://github.com/huanzhang12/ATLA_robust_RL

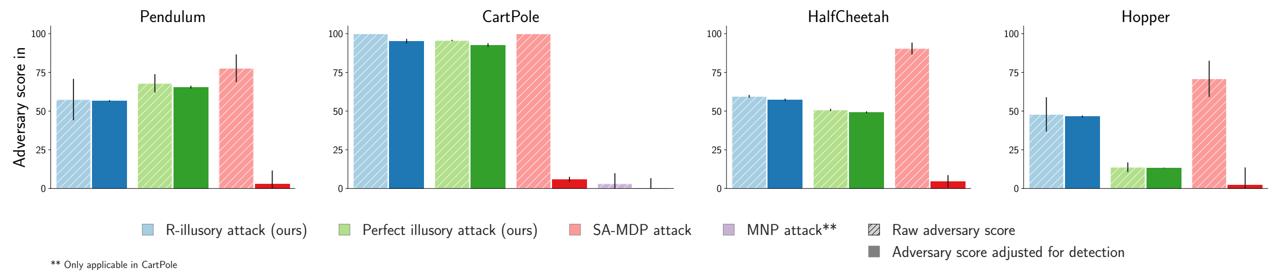

Figure 8: Results for $\beta = 0.05$. We display normalised adversary scores, indicating the reduction in the victim's reward, on the y-axis. Each plot shows results in different environments, with different adversarial attacks on the x-axis. We show both the raw adversary score, as well as the adversary score adjusted for detection rates of different adversarial attacks (see Figure 5). While the SA-MDP and MNP benchmark attacks achieve higher unadjusted scores, their high detection rates result in significantly lower adjusted scores. Note that MNP attacks perform significantly worse for $\beta = 0.05$, as compared to $\beta = 0.2$ (see Figure 4).

and training steps of the victim, but found that a ratio of one, i.e. equal training of both, yields the most stable results for co-training.

We implement the illusory adversary agent with SAC [25], where we likewise use the implementation given in Raffin et al. [59]. We initially ran a small study and investigated four different algorithms as possible implementations for the adversary agent, where we found that SAC yields best performance and training stability.

We train all adversarial attacks for three million environment steps. We implemented randomized smoothing as a standard defense against adversarial attacks on RL agents, as introduced in Kumar et al. [40]. We use the author's original implementation [6].

Computational overhead of R-illusory attacks. Note that there is no computational overhead of our method at test-time. We found in our experiments that the computational overhead during training of the adversarial attack scaled with the quality of the learned attack. In general, we found that the training wall-clock time for the R-illusory attacks results presented in Table 1 was about twice that of the SA-MDP attack (note that MNP attacks and perfect illusory attacks do not require training).

B.10.1. Results for perturbation budget $\beta = 0.05$

We show the remaining results for a perturbation budget of $\beta = 0.05$ in Figures 8 and 9, and in Table 2. Note that the corresponding Figures in the main paper are for a perturbation budget of $\beta = 0.2$.

[6] https://openreview.net/forum?id=mwdfai8NBrJ

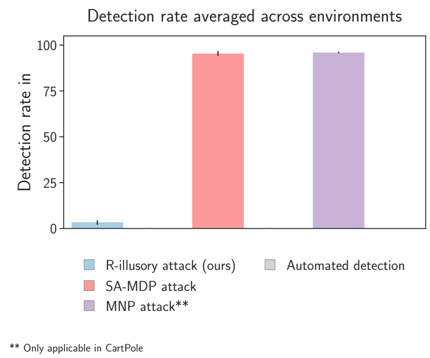

Figure 9: Results for $\beta = 0.05$. Different adversarial attacks are shown on the x-axis, with detection rates on the y-axis. We see that the automated reliably detector detects SA-MDP and MNP attacks, while R-illusory attacks are less likely to be detected. Perfect illusory attacks are excluded here as they are undetectable. Note that the study with human subjects did not contain examples with $\beta = 0.05$.

	Normalised adversary score in %			
Attack	no defence	smoothing	ATLA	ATLA abl.
MNP [40]	3 ± 7	64 ± 6	-	-
SA-MDP [91]	85 ± 7	50 ± 5	85 ± 4	83 ± 4
R-illusory attack (ours)	55 ± 8	47 ± 5	76 ± 6	70 ± 8
Perfect illusory attack (ours)	57 ± 6	63 ± 6	66 ± 3	65 ± 5

Table 2: Adversary scores and standard deviations averaged across environments for different defence methods and different attacks ($\beta = 0.05$). Defences decrease the adversary score, i.e., increase the victim reward across all classes and all attack algorithms.

B.10.2. Videos of all adversarial attacks

We provide a video summarising results in the supplementary material. Further, we provide videos for different seeds for all adversarial attacks in the supplementary material. The folders are named respectively. All videos were generated for a budget $\beta = 0.2$.

B.11. Human study

Study approval. Our study was approved by an independent ethics committee under reference xxxxx/xxxxx.

Setup. We performed a controlled study with $n = 10$ human participants. All participants were graduate-level university students. None had prior knowledge about the objective of the study. Participants participated voluntarily; we estimate the time needed per participant to be around 15 minutes. Participants were handed a slide show which contained all relevant information. This slide show is included in the supplementary material in the respective folder. We further add the sheet with ground truth labels for all video sequences.

After consenting to participate, participants were provided with the slide show and an online sheet to indicate their answers. The study was self-paced and done by the participants without further supervision. We further advised participants not to do the study if they would not feel comfortable at any point in time.

The first 18 videos in the slide show are 100-frame long sequences of the *Pendulum* environment, i.e. six each for either no adversarial attack, an R-illusory attack or an SA-MDP attack [91]. The subsequent 24 10-frame long sequences of the *CartPole* include the same attacks as for *Pendulum*, plus an additional six sequences under an MNP attack [40]. Note that MNP attacks cannot be applied to the *Pendulum* environment, as it has a continuous action space. For all attacks, the budget β was 0.2. Within both groups, the order of the videos shown was randomised. Subjects had to choose for each video between *correct* and *false*, where *false* indicates that subjects find that the system observed in the respective video is not the same as the one observed in the introductory video. We likewise refer to a video marked as *false* as a *suspicious* video. Table 5 includes all empirically estimated probabilities and standard deviations for classifying a segment from a given class as *false*.

Result analysis. We find that subjects reported one-third of unattacked sequences as *false*. Subjects classified sequences under an MNP or SA-MDP [90] attack as *false* with a probability higher than 80%, while categorizing illusory attacks as *false* with a probability of only 32.4%. We computed the double-sided z-statistics and were able to reject both the hypothesis that P(false | SA-MDP)=P(false | no attack) and the hypothesis that P(false | MNP)=P(false | no attack) for $\alpha = 0.05$, while the hypothesis that P(false | $\mathcal{E}$-illusory attack)=P(false | no attack) cannot be rejected. We conclude that subjects

Table 3: Full results table for all four environments

		Detection rate [%]		Victim reward under different defences			
attack	budget β	naive	ATLA[3]	none	smoothing	ATLA	ATLA abl.
Pendulum							
SA-MDP [91]	0.05	96.2± 0.01	95.4± 0.02	-797.2± 69.9	-408.4± 146.6	-757.2± 109.3	-722.2± 30.8
R-illusory attack (ours)		1.6± 0.01	2.1± 0.01	-638.8± 204.6	-387.8± 115.8	-634.4± 370.4	-634.9± 103.9
SA-MDP [91]	0.2	97.7± 0.01	93.8± 0.02	-1387.0± 119.0	-1188.3± 70.4	-1354.6± 107.1	-1428.3± 91.5
R-illusory attack (ours)		4.9± 0.01	4.7± 0.01	-1170.1± 67.5	-940.2± 91.6	-1020.4± 50.0	-1029.4± 106.7
Perfect illusory attack (ours)	1	3.6± 0.01	3.4± 0.01	-1204.8± 88.6	-1231.7± 25.3	-1284.5± 158.5	1228.6± 50.0
unattacked		3.2± 0.01	3.5± 0.01	-189.4			
CartPole							
MNP [40]		96.0± 0.01		485.0± 33.5	180.3± 33.6		
SA-MDP [91]	0.05	94.1± 0.02	94.5± 0.02	9.4± 0.2	122.5± 54.3	24.2± 7.3	16.8± 8.3
R-illusory attack (ours)		4.8± 0.01	4.6± 0.01	9.3± 0.1	165.4± 46.3	21.4± 6.0	45.4± 56.5
MNP [40]		95.2± 0.02		18.3± 20.8	20.8± 8.7		
SA-MDP [91]	0.2	99.7± 0.01	96.0± 0.01	9.3± 0.1	39.0± 10.7	9.2± 0.1	9.7± 0.6
R-illusory attack (ours)		3.7± 0.01	5.8± 0.02	9.0± 0.3	23.9± 3.3	9.6± 0.6	10.0± 1.20
Perfect illusory attack (ours)	1	3.1± 0.01	3.6± 0.01	30.1± 2.2	25.0± 1.6	21.9± 12.8	19.1± 4.6
unattacked		3.1± 0.01	3.3± 0.01	500.0			
HalfCheetah							
SA-MDP [91]	0.05	94.8± 0.02	94.9± 0.02	-1570.8± 177.4	101.3± 71.7	-570.2± 156.8	-625.3± 312.6
R-illusory attack (ours)		3.8± 0.01	4.7± 0.01	-149.1± 41.8	103.1± 44.8	-67.4± 47.3	-117.2± 2.3
SA-MDP [91]	0.2	97.1± 0.01	92.2± 0.02	-1643.8± 344.8	-36.8± 8.9	-1443.9± 313.8	-1200.7± 175.1
R-illusory attack (ours)		4.7± 0.01	4.3± 0.01	-178.9± 4.6	-31.0± 8.2	-64.7± 32.6	-35.5± 21.90
Perfect illusory attack (ours)	1	3.3± 0.01	3.4± 0.01	5.9± 36.8	-33.8± 4.7	153.0± 138.9	125.4± 107.5
unattacked		3.1± 0.01	3.5± 0.01	2594.6			
Hopper							
SA-MDP [91]	0.05	96.8± 0.01	96.4± 0.01	144.1± 265.4	488.5± 66.4	-205.7± 148.7	-124.0± 152.0
R-illusory attack (ours)		2.9± 0.01	6.2± 0.02	408.0± 127.2	489.5± 58.5	-79.4± 192.6	140.3± 69.4
SA-MDP [91]	0.2	95.6± 0.02	94.6± 0.02	-761.5± 127.4	255.1± 28.9	-274.8± 96.1	-502.6± 209.4
R-illusory attack (ours)		3.5± 0.01	3.9± 0.01	-157.5± 225.3	238.2± 24.8	-108.0± 106.6	-84.9± 176.4
Perfect illusory attack (ours)	1	3.2± 0.01	3.4± 0.01	679.2± 63.9	239.3± 32.9	59.5± 161.9	78.7± 165.6
unattacked		3.4± 0.01	3.0± 0.01	958.1			

were able to distinguish SA-MDP and MNP attacks from unattacked sequences while being unable to distinguish illusory attacks from unattacked sequences.

B.12. Reality feedback

We conclude our empirical investigations by exploring the importance of utilizing uncorrupted observation channels, which we refer to as *reality feedback*. We define reality feedback ζ as a part of the victim's observation $\mathcal{Z}$ in $\mathcal{E}'$ that cannot be corrupted by the adversary, *i.e.* , we assume that the victim's observations $\mathcal{Z} := \mathcal{Z}_0 \times \mathcal{Z}_\zeta$, where the adversary can modify $z^0 \in \mathcal{Z}_0$ but not $z^\zeta \in \mathcal{Z}_\zeta$. We establish two reality feedback scenarios for *CartPole*: one where the cart observation is unattacked, and one where the observation of the pole is unattacked. We find that robustifying the victim agent through adversarial training allows victim policies to use reality feedback effectively at test-time. Our results further suggest that having access to reality feedback channels allows for significant robustification if those channels are sufficiently *informative*. In the scenarios studied, we found that having access to an unattacked observation of the pole is more valuable than having access to an unattacked observation of the cart. See App. B.12 for details.

Setup. We evaluate the importance of realism feedback in the *CartPole* environment by investigating two possible scenarios. Note that the observation in *CartPole* is given as a four-dimensional vector of the pole angle and angular velocity,

Table 4: Reward achieved by victim for different reality feedback scenarios.

	Victim agent	
Reality feedback	naive	ATLA abl.
Pole	9.84 ± 0.1	182.44 ± 36.9
Cart	8.83 ± 0.3	15.54 ± 6.6

Table 5: Results from our study with human participants.

	Environment		
	both	Pendulum	CartPole
P(false | no attack)	34.2 ± 11.4	31.5 ± 10.5	37.0 ± 12.3
P(false | SA-MDP)	81.4 ± 27.2	96.3 ± 32.1	66.7 ± 22.2
P(false | R-illusory attack)	32.4 ± 10.8	37.0 ± 12.3	27.7 ± 9.3
P(false | MNP)	83.3 ± 27.8		83.3 ± 27.8

as well as cart position and velocity. In the first test scenario, the victim correctly observes the pole, while the adversary can attack the observation of the cart; the second scenario is vice versa. We investigate two test cases for each scenario: First, attacking a naive victim, and second, attacking an agent pretrained with co-training.

Results and discussion. Table 4 shows that the reward achieved by the victim is generally higher when pretrained with co-training. We hypothesize that this pretraining enables the agent to learn how to utilize the reality feedback effectively. The achieved victim performance when reality feedback contains information about the *pole* is more than 10 times larger than when containing information on the *cart* instead. This seems intuitive, as the observation of the pole appears much more useful for the task of stabilizing the pole, and underlines the importance of equipping agents with strong reality feedback channels.