Camouflage Adversarial Attacks on Multi-agent Reinforcement Learning Systems

Ziqing Lu, Guanlin Liu, Lifeng Lai, Weiyu Xu

***Abstract*—The multiple agent reinforcement learning systems (MARL) based on the Markov Game (MG) have emerged in many critical applications. To improve the robustness/defense of MARL systems against adversaries, studying various adversarial attacks on reinforcement learning systems is very important. Previous works on adversarial attacks considered some possible features to attack in MDP, such as the action poisoning attacks, the reward poisoning attacks, and the state perception attacks. In this paper, we propose a brand-new form of attack called the camouflage attack in the MARL systems. In the camouflage attack, the attackers change the appearances of some objects in the environment but without changing the actual objects; and the camouflaged appearances may look the same to all the targeted recipient (victim) agents. The camouflaged appearances can mislead the recipient agents to follow misguided policies. We evaluate the effect of camouflage attacks in two different scenarios: Camouflage attacks were performed during the learning (training-time attacks) and were performed during the test of agents' policies (test-time attacks). Our numerical and theoretical results show that camouflage attacks can rival the more conventional, but likely more difficult state perception attacks, by comparing their effect on reducing agents' global benefits. We also investigated cost-constrained camouflage attacks and showed how cost budgets affect attack performance numerically.**

***Index Terms*—Muti-agent Reinforcement learning, Adversarial attacks, Markov decision process**

I. INTRODUCTION

SINGLE-AGENT and multiple-agent reinforcement learning (RL) algorithms are used in many safety or security-related applications, such as autonomous driving [1], financial decisions [2], recommendation systems [3], wireless communication [4], and also in drones' and robots' algorithms [5]. It is thus essential to develop trustworthy systems before their real-world deployment. Studying the potential adversarial attacks on RL systems and evaluating the worst-case performances of RL agents under these attacks can help us limit the damage imposed by adversarial parties, defend against adversarial attacks, and therefore build more robust and secure RL systems.

Adversarial attacks and defenses against these attacks for single-agent RL systems have been relatively well studied so far [6]–[17], but adversarial attacks on multi-agent learning are still not well understood. The challenges of MARL such as scalability and non-stationarity potentially make it harder to find robust solutions under environmental perturbations. A series of model-based algorithms aim to find equilibrium solutions of MARL [18]–[20]. However, none of them considered adversary settings in their problem formulations. [21] considered the white and black box adversarial attacks on MARL and used new concepts like attack loss and cost functions to evaluate the efficacy of adversarial attacks on MARL systems.

In terms of the types of adversarial attacks on MARL, most proposed adversarial attacks only consider recipient (victim) agents' properties to attack, for example, the action poisoning attacks, the reward poisoning attacks, the state poisoning attacks, the environmental attacks, or the mixed attacks [6], [21]–[28]. These attacks either directly change the features of agents, i.e., actions, rewards, or states of the MDP, or perturb the interactions between the agents' actions and the environments. In [29], [30], the authors proposed a form of state perception (observation) attack in deep reinforcement learning, in which attackers confuse agents with delusional states instead of changing their actual states during the game. In [31], the authors addressed the state perception attacks with cost constraints in a multi-agent system.

There is a close relationship between the study of adversarial attacks and robust reinforcement learning. Robust reinforcement learning aims to learn a policy that optimizes the worst-case performance within a prescribed uncertainty of transition kernels. In adversarial reinforcement learning, agents are trained in the presence of an adversary that applies disturbance to the environment. There are some studies aiming to develop robust policies by addressing challenges in the adversarial framework. For example, [29] focuses on improving the robustness of deep reinforcement learning by proposing alternating training with learned adversaries in a state-adversarial MDP. In [32], the authors formulate robust RL as a constrained minimax game between the RL agent and an adversary that controls uncertainties in the environment as adversarial disturbances.

In this paper, we propose a new form of adversarial attack on MARL system: the camouflage attack. During the camouflage attack, instead of directly changing recipients' properties, the attackers change the appearances of some objects they can control or even the appearances of attackers themselves. After the camouflage attack, all the recipient agents potentially observe the same camouflaged objects' features so that they are misled to misguided decisions in the MG. The camouflage attacks are different from the state perception attacks in two ways: 1) the camouflage attack does not directly change the measurements of each recipient agents, but instead change the appearances of the objects the attackers can control thus changing the measurements of the victim agents indirectly; 2) in the camouflage attack, the perceptions of different recipient agents cannot be freely manipulated as in state perception

[1] Parts of results have been published in the conference ISIT 2024.

attacks: the confusions of the recipient agents come from observing the same camouflaged objects and thus are the same or correlated. In addition, in camouflage attacks, the underlying true states of the camouflaged objects are not changed, and what are changed are only the appearances of the camouflaged objects. For example, camouflaged robot examples include stealthy invisible airplanes that can evade the detection of regular radars: they are actually in the air but are "camouflaged" to be invisible. As another example, one can camouflage unimportant objects into fake "valuable" targets so that enemy robots spend precious resources on attacking these fake targets.

There are barely any studies of camouflage attacks from the perspective of RL. Even though the terminology "camouflage" is used in [33], it discussed essentially state perception attack for single-victim-agent dynamical systems. For non-dynamical systems, some works discussed improving the detection of camouflaged attacks in deep learning models [34], [35].

MARL builds on the Markov game, which combines the Markov decision process and game theory. The solution to MARL is usually considered an equilibrium policy (Nash, coarse-correlated, correlated) at which no agent can unilaterally improve its expected reward. In the context of this paper, agents aim to maximize their long-term anticipated reward in the shared environment while adapting to other agents' strategies. The solution to MARL in this paper is a set of policies that jointly maximize global benefit.

Parts of test-time camouflage attacks were published in [36], but we significantly extended our results by introducing two brand-new training-time camouflage attacks, and provided new extensive theoretical performance analysis. The paper is organized as follows: Section II introduces the test-time and training-time camouflage attack model. Section III analyzes the camouflage attacks, showing they perform comparably to the more costly state perception attacks. Section IV provides numerical evaluations, which demonstrates a significant drop in expected reward and supports the theory in Section III.

II. Problem Formulation

In the considered MDP environment, all the agents are divided into two opposite groups, the attacker group M, and the recipient (victim) agent group N, with $|M| = m$, $|N| = n$.

The 5-element tuple can describe the finite MDP environment for the recipient agents: $(\{\mathcal{S}_i\}_{i=1}^n, \{\mathcal{A}_i\}_{i=1}^n, \{P^i\}_{i=1}^n, \{R_i\}_{i=1}^n, T)$, where T is the finite number of time steps, $\mathcal{S}_i$ is the state space of the i-th recipient agent with $|\mathcal{S}_i| = S_i$, $\mathcal{A}_i$ is the action space for the i-th recipient agent with $|\mathcal{A}_i| = A_i$. We let $P^i_{t\to t+1} : \mathcal{S}_i \times \mathcal{A}_i \times \mathcal{S}_i \to [0, 1]$ be the i-th agent's transition probability kernel between time indices t and $t+1$ with $t = 0, 1, \dots T-1$. We use $R_{i,t} : \mathcal{S}_i \times \mathcal{A}_i \to \mathbb{R}$ to represent the reward function of the i-th recipient agent at time index t, with $t = 1, 2, \dots, T$. The reward function settings depend on the exact problem. Our time index t starts from 0 to T. We refer to **time step** t as the time interval starting from time index $t-1$ and ending at time index t, where $1 \le t \le T$.

We let $a_{i,t}$ denote the action the i-th agent takes at time index t and denote $a^*_{i,t}$ the optimal action of the i-th agent at time step t. The joint action of all n agents at time index t is $\mathbf{a}_t := (a_{1,t}, a_{2,t}, \dots, a_{n,t})$. We denote $\pi^*_{i,t}$ the optimal policy of the i-th recipient agent at time index t, and $\pi^*_{i,t}(s_{i,t}) = a^*_{i,t}$. The deterministic joint optimal policy at time index t is $\boldsymbol{\pi}^*_t = \{\pi^*_{i,t}\}_{i=1}^n$, and we write $\pi^*_i = \{\pi^*_{i,t}\}_{t=0}^{T-1}$.

We use $V_{i,t}^{\pi_i} : \mathcal{S}_i \to \mathbb{R}$ to denote the value function of agent i at time index t under guidance of the individual policy π_i, where $V_{i,t}^{\pi_i}(s_i) = \mathbb{E}_{\pi_i}[\sum_{k=0}^t R_{i,k}(s_{i,k}, a_{i,k}) | s_{i,0} = s_i]$. The Q-function $Q_{i,t}^{\pi_i} : \mathcal{S}_i \times \mathcal{A} \to \mathbb{R}$ of agent i at time index t with a state-action pair (s_i, a_i) is defined to be $Q_{i,t}^{\pi_i}(s_i, a_i) = \mathbb{E}_{\pi_i}[\sum_{k=0}^t R_{i,k}(s_{i,k}, a_{i,k}) | s_{i,0} = s_i, a_{i,0} = a_i]$. The relation between Q-value and V-value is $Q_{i,t}^{\pi_i}(s_i, a_i) = R_{i,t}(s_i, a_i) + \sum_{a_i'} \pi_i(a_i'|s_i') \sum_{s_i'} P^i_{t\to t+1}(s_i'|s_i, a_i) V_{i,t+1}^{\pi_i}(s_i')$.

In our setting to introduce camouflage attacks, for the sake of simplification, the overall state space is represented by a factored representation $\{\mathcal{S}_i\}_{i=1}^n$. This happens when each agent acts independently, its action only affects its own state but the camouflage of the common environment objects affects the states of every agent. In a more general setting, the actions of agents can affect each other's states jointly. In section IV-E, we considered a scenario where all agent share a state space $\mathcal{S}$ and their joint action $\boldsymbol{a}_t$ decides the proceedings of MG.

We assume a white-box case such that the attackers can monitor the underlying MARL algorithms of the recipient agents, and therefore attackers know optimal policies $\pi^*_{i,t}$ of every recipient agent i at time index t. However, the recipient agents are unaware of the existence of attackers or their attacks. The m attackers perform camouflage attack by disturbing recipient agents' observations of their true states. For a recipient agent i at time index t, we let $s_{a,t,i}$ denote the true state the agent i is actually in and let $s_{d,t,i}$ denote the delusional state that the agent i thinks it is in.

The recipient agents are selfish in the game, aiming to maximize their own rewards obtained during the T time steps. The attackers aim to minimize the recipient agents' total expected rewards. We discuss test-time and training-time camouflage attacks separately in this paper, based on when the camouflage attacks occur.

III. Test-time Camouflage Attacks

In test-time camouflage attacks, recipient agents have been trained in the ground truth finite MDP without any attacks, and they are aware of their optimal policies, $\pi^*_{i,t}$ after the training. The attackers only perform camouflage attacks during the test of the optimal policies $\pi^*_{i,t}$ at every time step t.

There are two phases of play during one time step t in the test. In the first phase, from time index $t-1$ to $t-0.5$, the attackers attack to make each recipient agent i ($1 \le i \le n$) think it is in a delusional state $s_{d,t-0.5,i}$. In the 2nd phase, after the attack, from the time index $t-0.5$ to t, each recipient agent i moves to $s_{a,t,i}$ according to its policy at time step t, $\pi_{i,t}(s_{d,t-0.5,i}) = a_{i,t}$, in which $s_{d,t-0.5,i}$ is agent i's delusional state at time index $t-0.5$, and obtains its corresponding reward $r_{i,t} = R_{i,t}(s_{a,t-1,i}, a_{i,t}, s_{a,t,i})$, as described in Figure 1. We are interested in finding the optimal attack strategy of attackers for each time step t ($1 \le t \le T$).

Test-time camouflage attack: The m attackers can change the appearance of some objects that they control during the

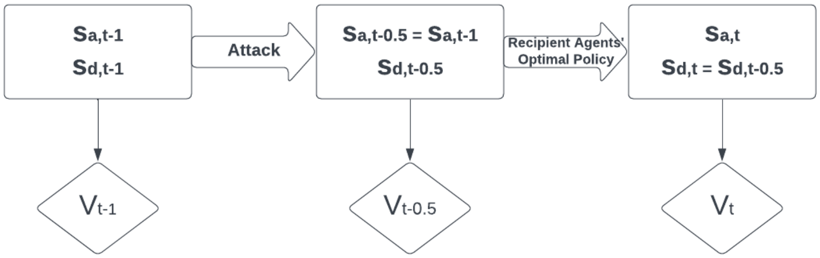

Figure 1: Transition between states

test at every time step t. Mathematically, suppose that we have a random variable $X_{t-0.5}$ which represents the true status of an object at time index $(t-0.5)$. If it is not camouflaged, the appearance, denoted by $Y_{t-0.5}$, of this object is just equal to $X_{t-0.5}$, namely $Y_{t-0.5} = X_{t-0.5}$. The camouflage attack changes the appearance $Y_{t-0.5}$ to some other value, for example, $Y_{t-0.5} = g(X_{t-0.5})$ where $g(\cdot)$ is a camouflage function. Then the observation of $X_{t-0.5}$ at time index $(t-0.5)$ from the perspective of agent i is given by $s_{d,t-0.5,i} = h_i(g(X_{t-0.5})) = h_i(Y_{t-0.5})$, where $Y_{t-0.5}$ is the changed appearance of the object, and h_i is the observation function of recipient agent i (h_i can be a function giving random outcomes, for example, due to noises).

These camouflaged objects will fool the recipient agents into a distribution of delusional states $s_{d,t-0.5,i}$ for each agent i. Different from the state perception attacks, these delusional states $s_{d,t-0.5,i}$ have to be correlated or even the same across different recipient agents: this is because camouflage attacks make the recipient agents observe the same camouflaged objects. In state perception attacks [29], the attackers can instead fool different recipients into very different delusional states.

Cost constrained test-time camouflage attacks: In practice, sometimes the attackers have attack budgets that must be spent by the end of each time step. For example, the resources used by attackers are provided by constantly-energy-harvesting systems over time steps, and the budget for each time step is constrained by the battery volume. We call this scenario an "instant cost constrained case". Within each time step t (namely between time index $t-1$ and t), all attackers share a budget B, and this budget B can only be spent during that single time step: the leftover resources cannot be carried over to the next time step $t+1$ or there is no need to carry over the leftover resources to the next time step because of budget refill. Once we get to time step $t+1$, the shared budget B will be refilled (say, to B). We would like to find out how to optimally allocate the total resources to each attacker j for performing the camouflage attack in each time step while satisfying the instant cost constraint and minimizing the recipient agents' total rewards. To simplify our presentations, we consider the budgets are used to camouflage the attackers themselves.

We design an integration of between-step dynamic programming and within-step static constrained optimizations to compute the optimal attack strategy. During each time step t, for each possible actual state vector $\mathbf{s}_{a,t-1}$, we use a static standalone optimization program to determine the optimal allocation of attackers' budgets on camouflages. Between different time steps, we use dynamic programming to account for the state transitions and expected rewards.

We work backward from time index $t = T$ and initialize value function $V_T^{\star}(\sigma_T) = 0$ for each dynamic programming state (DPS) σ_T (a DPS state includes all recipient and attacker agents' actual conditions, and also the conditions of camouflaged objects), and the subscript represents time index. Suppose that we have already computed $V_{t+0.5}^{\star}(\sigma_{t+0.5})$ for every DPS state $\sigma_{t+0.5}$. We then proceed to compute the optimal attack policy during time step t (essentially from t to $t+0.5$) and also $V_t^{\star}(\sigma_t)$ for every DPS state σ_t. During time step t, we let $b_j \in \mathbb{R}$ be the amount of resources attacker j spends on its camouflage attack. The constraints on b_j are such that the total spending of all attackers cannot exceed B. To represent formulas concisely, we stack the b_j's to form a m-dimensional vector $\mathbf{b}$ called the attack allocation vector. Under the attack allocation vector $\mathbf{b}$, we denote the probability that the recipient agents' state will transit to $\sigma_{t+0.5}$ as $P(\mathbf{b}, \mathbf{s}_{a,t}, \sigma_{t+0.5})$, where $\mathbf{s}_{a,t}$ is the true states of all the recipient agents at time index t. This probability must be between 0 and 1. Based on the principles of dynamic programming, we want to optimize b_j's to minimize the total expected rewards received by the agents from time step t to T. Thus, under a specific true state vector $\mathbf{s}_{a,t}$, the objective function for attackers to minimize is the expected total reward of all the recipient agents from step t onward to step T.

1) Within-step static constrained optimization problem: Suppose that the DPS has Q possible values at time index $t+0.5$ conditioned on the true states are $\mathbf{s}_{a,t}$, the optimal attack under the "instant cost constrained" case at a single time step t can be formulated as the following within-step static constrained optimization problem:

$$\begin{aligned} \min \sum_{k=1}^{Q} & P(\mathbf{b}, \mathbf{s}_{a,t}, \sigma_{t+0.5}^{k}) V_{t+0.5}^{\star}(\sigma_{t+0.5}^{k}) \qquad (1) \\ \text{subject to} \sum_{j=1}^{m} & b_j \leq B, \\ & P(\mathbf{b}, \mathbf{s}_{a,t}, \sigma_{t+0.5}^{k}) \leq 1, \ \forall k \\ & -P(\mathbf{b}, \mathbf{s}_{a,t}, \sigma_{t+0.5}^{k}) \leq 0, \ \forall k \\ & b_j \geq 0, \ j = 1, \ldots, m, \end{aligned}$$

where $\sigma_{t+0.5}^{k}$ is the k-th DPS at time index $t+0.5$.

Depending on the physical nature of the attacks, we can model the probability $P(\mathbf{b}, \mathbf{s}_{a,t}, \sigma_{t+0.5}^{k})$ as a function of $\mathbf{b}$. In one particular model considered in the paper, for each attacker j, the probability that it can change the appearance of the object it controls is $\min\{b_j/C_t(x_j, y_j), 1\}$, where $C_t(x_j, y_j)$ are constants representing how hard it is for the attacker j to camouflage the appearance of x_j as y_j. In our numerical results, we take $C_t(x_j, y_j) = d(s_{a,j}^{\dagger}, s_{d,j}^{\dagger}) + \epsilon$ where ϵ is a positive constant and $d(s_{a,j}^{\dagger}, s_{d,j}^{\dagger})$ is the distance between the real position of the attacker j and the target camouflaged position the attacker j chooses. Namely, if we assign more budget to attacker j, and if the target camouflage position is closer to its actual position, it is more likely that attacker j can change the objects to the targeted appearances.

2) Between-step dynamic programming: After solving (1), we take the optimal value of its objective function as $V_t^{\star}(\sigma)$, using which we continue to calculate $V_{t-0.5}^{\star}(\cdot)$ as follows. For each DPS $\sigma_{t-0.5}$, we update $V_{t-0.5}^{\star}(\sigma_{t-0.5})$ as

$$\sum_{\sigma_t} P(\sigma_t|\sigma_{t-0.5}, \mathbf{a}_{t-0.5}^{\star})(V_t^{\star}(\sigma_t) + R(\sigma_{t-0.5}, \sigma_t)).$$

After updating $V_{t-0.5}^{\star}(\cdot)$, again we will use another static optimization formulation in Section III-1 to calculate $V_{t-1}^{\star}(\cdot)$. In this way, we perform this static optimization-dynamic programming cycle recursively until we calculate all the $V_t^{\star}(\cdot)$ backward from $t = T$ until $t = 0$.

IV. Training-time Camouflage Attacks

In training-time camouflage attacks, however, the attackers can change the appearance of objects during the learning, which leads to recipient agents' confused observations of transition kernels and confused observations of reward functions. From the perspective of any agent i, let $P^{d,i} = \{P_{t\to t+1}^{d,i}\}_{t=0}^{T-1}$ denote the perceived transition kernels during the learning, and let $R_i^d = \{R_{i,t}^d\}_{t=1}^{T}$ denote the perceived reward functions during the learning. We assume that in practice, agents are trained over a sufficiently large number of episodes K so that each agent learns the perceived $P^{d,i}$, R_i^d instead of the ground-truth P^i, R_i, and therefore every agent i learns a delusional "optimal" policy $\pi_{i,t}^{d,*}$ after the training. The attackers only perform camouflage attacks during the training of agents. In the test of the delusional "optimal" policies $\pi_{i,t}^{d,*}$, no attack occurs.

Training-time camouflage attacks: The m attackers change the appearance of agents' states according to the camouflage function $g(s_{d,i}|s_{a,i}) : \mathcal{S}_i \times \mathcal{S}_i \to [0,1]$, such that $g(s_{d,i}|s_{a,i}) = a$, $a \in [0,1]$. The camouflage function g means that the attackers change the appearance of the true state $s_{a,i}$ into the delusional state $s_{d,i}$ with probability a. Moreover, $\sum_{s_{d,i}\in\mathcal{S}_i} g(s_{d,i}|s_{a,i}) = 1$. The exact form of g depends on the type of training-time camouflage attacks.

To emphasize the idea of camouflage attacks, the camouflage function g needs to be correlated or even exactly the same across different recipient agent i, due to the fact that camouflage attacks make the recipient agents observe the same camouflaged objects.

Perceived transition kernels and perceived reward functions of individual recipient agents are affected due to the change of appearance in the learning. Let $P^i(s'|s,a) : \mathcal{S}_i \times \mathcal{A}_i \times \mathcal{S}_i \to [0,1]$ be the ground-truth transition kernel of a given state-action pair $(s,a) \in (\mathcal{S}_i, \mathcal{A}_i)$ and a given next state $s' \in \mathcal{S}$ of the i-th agent's. If the agent is not camouflaged, its perceived transition kernels and reward functions are the same as the ground-truth kernels and reward functions: $P^{d,i} = P^i$, $R_i^d = R_i$. The camouflage attack changes the perception $P^{d,i}$ to some other value that is other than the true P^i, and so for the perceived reward functions. We discuss how to find $P^{d,i}$ and R_i^d at every time step precisely under two different training-time camouflage attacks: the *one-time camouflage attack*, and the *permutation camouflage attack*.

A. One-time camouflage attack

Settings: We discuss the one-time camouflage attack from the perspective of an individual recipient agent i. We assume that the agent i follows a static stochastic policy $\pi_{i,t}(a_i|s_i) = 1/A_i$, $\forall a_i \in \mathcal{A}_i$, $\forall s_i \in \mathcal{S}_i$, $\forall t$ in the learning. During the learning, the recipient i learns a series of transition kernels $\{P_{t\to t+1}^{d,i}\}_{t=0}^{T-1}$, camouflaged reward functions $\{R_{i,t}^d\}_{t=1}^{T}$ and uses dynamic programming to solve for the delusional "optimal" policy $\{\pi_{i,t}^{d,*}\}_{t=0}^{T-1}$.

The attackers launch the camouflage attack **once** at a time index t^* in the whole learning, $0 \leq t^* \leq T-1$. This camouflage attack fools the recipient agent i on its actual state $s_{a,t^*,i} \in \mathcal{S}_i$ with a stochastic delusional state $s_{d,t^*,i} \in \mathcal{S}_i$ at the time index t^*. At other time indices t such that $t \neq t^*$, the agent i can correctly perceive its true state $s_{a,t,i} \in \mathcal{S}_i$. The camouflage function $g(s_{d,t^*,i}|s_{a,t^*,i})$ gives the camouflage probability a of deceiving the agent with the delusional state $s_{d,t^*,i}$ given its actual state $s_{a,t^*,i}$. In the one-time stochastic camouflage attack, we use $x_{s_{a,t^*,i}\to s_{d,t^*,i}}$ to denote a. The parameters $\{x_{s_{a,t^*,i}\to s_{d,t^*,i}}\}_{s_{a,t^*,i}, s_{d,t^*,i}\in\mathcal{S}_i}$ are subject to the conditions that:

$$\begin{cases} \sum_{s_{d,t^*,i}} x_{s_{a,t^*,i}\to s_{d,t^*,i}} = 1, \ \forall s_{a,t^*,i} \\ 0 \leq x_{s_{a,t^*,i}\to s_{d,t^*,i}} \leq 1. \end{cases} \tag{2}$$

Learning: For every agent i, denote the actual state distribution as $\{p_t(s_{a,t,i})\}_{t=0}^{T}$. We assume that the distribution of the initial true state $\{s_{a,0,i}\}$ is uniform, i.e., $p_0(s_{a,0,i}) = 1/S_i$,for any initial state $s_{a,0,i}$. There is an iterative relation between the previous state distribution $p_t(s_{a,t,i})$ and the next state distribution $p_{t+1}(s_{a,t+1,i})$ such that

$$\begin{aligned} p_{t+1}(s_{a,t+1,i}) &= \sum_{s_{a,t,i}} \sum_{a_i} P_{t\to t+1}^{i}(s_{a,t+1,i}|s_{a,t,i}, a_i)\pi_{i,t}(a_i|s_{a,t,i})p_t(s_{a,t,i}) \\ &= \sum_{s_{a,t,i}} \sum_{a_i} P_{t\to t+1}^{i}(s_{a,t+1,i}|s_{a,t,i}, a_i)1/A_i \ p_t(s_{a,t,i}), \end{aligned}$$

where $P_{t\to t+1}^{i}(s_{a,t+1,i}|s_{a,t,i}, a_i)$ is the ground-truth transition probability.

Due to the one-time attack, the agents learn a series of in-homogeneous transition kernels throughout time steps. For every agent i, we use $P_{t\to t+1}^{d,i}(s_{t+1,i}|s_{t,i}, a_i)$ to denote the perceived transition probability of moving from the current state $s_{t,i}$ with action a_i, to the next state $s_{t+1,i}$. Since the attackers strike only once at time index $t = t^*$, any agent i observes the ground-truth transition probability for $t \neq t^* - 1, t^*$.

During the time step between $t^* - 1$ and t^*, the agent i can observe the true state $s_{a,t^*-1,i}$ at time index $t^* - 1$. With action a_i, the true distribution of the next actual state $s_{a,t^*,i}$ follows the ground-truth transition probability $P^i_{t^*-1 \to t^*}(s_{a,t^*,i}|s_{a,t^*-1,i}, a_i)$. However, before the agent i observes its next true state $s_{a,t^*,i}$, the attackers launch a camouflage attack at t^*, causing the agent to believe that it's in the delusional state $s_{d,t^*,i}$. The probability of changing the state perception from $s_{a,t^*,i}$ to $s_{d,t^*,i}$ follows the camouflage probabilities $x_{s_{a,t^*,i} \to s_{d,t^*,i}}$. Therefore, the agent i's perceived transition kernel between $t^* - 1$ and t^* is:

$$\begin{aligned} &P^{d,i}_{t^*-1 \to t^*}(s_{t^*,i}|s_{t^*-1,i}, a_i) \\ &= \text{Prob}(s_{d,t^*,i} = s_{t^*,i}|s_{t^*-1,i} = s_{a^*,t-1,i}, a_i), \ \forall a_i. \end{aligned}$$

This perceived transition kernel is the probability of moving from the true state $s_{a,t^*-1,i}$, following action a_i, to the next delusional state $s_{d,t^*,i}$, which can be computed as (3).

$$P^{d,i}_{t^*-1 \to t^*}(s_{t^*,i}|s_{t^*-1,i}, a_i) = \sum_{s_{a,t^*,i}} P^i_{t^*-1 \to t^*}(s_{a,t^*,i}|s_{a,t^*-1,i}, a_i) x_{s_{a,t^*,i} \to s_{d,t^*,i}}, \tag{3}$$

$$P^{d,i}_{t^* \to t^*+1}(s_{t^*+1,i}|s_{t^*,i}, a_i) = \frac{\sum\limits_{s_{a,t^*,i}} P^i_{t^* \to t^*+1}(s_{a,t^*+1,i}|s_{a,t^*,i}, a_i) x_{s_{a,t^*,i} \to s_{d,t^*,i}} \pi_{i,t}(a_i|s_{d,t^*,i}) p_{t^*}(s_{a,t^*,i})}{\sum\limits_{s_{a,t^*,i}} x_{s_{a,t^*,i} \to s_{d,t^*,i}} \pi_{i,t}(a_i|s_{d,t^*,i}) p_{t^*}(s_{a,t^*,i})}, \tag{4}$$

$$R^d_{i,t^*}(s_{t^*,i}, a_i) = \sum_{s_{a,t^*,i}} \frac{x_{s_{a,t^*,i} \to s_{d,t^*,i}} p_{t^*}(s_{a,t^*,i}) R_{i,t^*}(s_{a,t^*,i}, a_i)}{\sum_{s_{a,t^*,i}} x_{s_{a,t^*,i} \to s_{d,t^*,i}} p_{t^*}(s_{a,t^*,i})}. \tag{5}$$

During the time step between t^* and $t^* + 1$, agent i thinks it is in a delusional state $s_{d,t^*,i}$ but it is actually in $s_{a,t^*,i}$ at t^*. Then, the recipient agent takes the action a_i, and observes the next true state $s_{a,t^*+1,i}$ at $t^* + 1$ since there are no more attacks. The agent perceives another camouflaged transition kernel during the time step between t^* and $t^* + 1$:

$$P^{d,i}_{t^* \to t^*+1}(s_{t^*+1,i}|s_{t^*,i}, a_i) = \text{Prob}(s_{t^*+1,i}|s_{d,t^*,i} = s_{t^*,i}, a_i), \ \forall a_i.$$

It is the probability/ of moving from the delusional state $s_{d,t^*,i}$ with action a_i to the next actual state $s_{a,t^*+1,i}$. We apply conditional probability to solve for it:

$$P^{d,i}_{t^* \to t^*+1}(s_{t^*+1,i}|s_{t^*,i}, a_i) = \frac{p(s_{a,t^*+1,i}, s_{t^*,i} = s_{d,t^*,i}, a_i)}{p(s_{t^*,i} = s_{d,t^*,i}, a_i)}.$$

Because

$$\begin{aligned} &p(s_{a,t^*+1,i}, s_{t^*,i} = s_{d,t^*,i}, a_i) \\ &= \sum_{s_{a,t^*,i}} p(s_{a,t^*+1,i}, s_{t^*,i} = s_{d,t^*,i}, a_i, s_{a,t^*,i}), \end{aligned}$$

and this can be extended to

$$\frac{\sum\limits_{s_{a,t^*,i}} p(s_{a,t^*+1,i}|s_{a,t^*,i}, a_i) p(s_{d,t^*,i}|s_{a,t^*,i}) p(a_i|s_{d,t^*,i}) p(s_{a,t^*,i})}{\{\sum\limits_{s_{a,t^*,i}} p(s_{d,t^*,i} = s_{t^*,i}|s_{a,t^*,i}) p(s_{a,t^*,i})\} p(a_i|s_{d,t^*,i} = s_{t^*,i})},$$

where $p(\cdot)$ denotes the true probability. In the above formula, $p(s_{a,t^*+1,i}|s_{a,t^*,i}, a_i)$ is the ground-truth transitional probability, $p(s_{d,t^*,i}|s_{a,t^*,i})$ is the camouflage probability, $p(a_i|s_{d,t^*,i})$ is the static policy, and $p(s_{a,t^*,i})$ is the true state distribution at t^*. By replacing these variables with notations we defined before, we get (4). Notice that all the parameters in (4) are known based on the settings of MDP and the one-time camouflage attack.

The one-time camouflage attack at t^* also affects the agent i's perception of the reward functions. Let $R^d_{i,t}(s_{i,t}, a_i)$ represent the perceived reward function at time index t in the learning. Since attackers only launch one camouflage attack at the time index t^*, $R^d_{i,t}(s_{i,t}, a_i) = R_{i,t}(s_{a,i,t}, a_i), \ \forall a_i,$ for $t \neq t^*$. However, at time index t^*, the agent i learns a camouflaged reward function such that $R^d_{i,t^*}(s_{i,t^*}, a_i) = R_{i,t^*}(s_{d,t^*i}, a_i), \ \forall a_i$. The actual reward value returned from the environment still depends on the actual state $s_{a,t^*,i}$, but the agent i thinks the returned reward value is resulted from the delusional state $s_{d,t^*,i}$. The exact form of camouflaged reward function is as the following:

$$\begin{aligned} R_{i,t^*}(s_{d,t^*,i}, a_i) &= \sum_{s_{a,t^*,i}} p(s_{a,t^*,i}|s_{d,t^*,i}) R_{i,t^*}(s_{a,t^*,i}, a_i) \\ &= \sum_{s_{a,t^*,i}} \frac{p(s_{d,t^*,i}|s_{a,t^*,i}) p_{t^*}(s_{a,t^*,i}) R_{i,t^*}(s_{a,t^*,i}, a_i)}{p_{t^*}(s_{d,t^*,i})}, \end{aligned}$$

where $p_{t^*}(s_{d,t^*,i}) = \sum_{s_{a,t^*,i}} p_{t^*}(s_{a,t^*,i}) p(s_{d,t^*,i}|s_{a,t^*,i})$. Therefore, the perceived reward at time index t^* is (5).

Under the one-time camouflage attack, We summarize the perceived transition kernels $\{P^{d,i}_{t \to t+1}\}_{t=0}^{T-1}$ and perceived reward functions $\{R^d_{i,t}\}_{t=1}^{T}$ as the following:

$$P^{i,d}_{t \to t+1}(s_{i,t+1}|s_{i,t}, a_i) = \begin{cases} (3), \ for \ t = t^* - 1 \\ (4), \ for \ t = t^* \\ \text{Ground-truth } P^i_{t \to t+1}, \ otherwise \end{cases}$$

$$R^d_{i,t}(s_{i,t}, a_i) = \begin{cases} (5), \ for \ t = t^* \\ \text{Ground-truth } R_{i,t}, \ otherwise \end{cases}$$

Notice that with the camouflage attack, for every agent i, the camouflage probabilities $\{x_{s_{a,t^*} \to s_{d,t^*}}\}$ are the same across all agents.

B. Permutation camouflage attack

Settings: The settings of the recipient agents are the same as those in the one-time camouflage attack. However, in this case the attackers launch the permutation camouflage attack at every time step t during the training of recipient agents. Let the permutation function be $\sigma : \mathcal{S}_i \to \mathcal{S}_i$. In this case, the camouflage function satisfies that

$$g(s_{d,i}|s_{a,i}) = \begin{cases} 1, \text{ if } \sigma(s_{a,i}) = s_{d,i} \\ 0, \ otherwise \end{cases}$$

Let $\Sigma_i = [a_{pq}] \in \{0,1\}^{S_i \times S_i},\ p,q \in \{1,2,\ldots,S_i\}$, denote the permutation matrix and Σ_i is subject to the following conditions: $\sum_{p=1}^{S_i} a_{pq} = 1, \forall q$, and $\sum_{q=1}^{S_i} a_{pq} = 1, \forall p$.

We apply the same permutation attack Σ_i at every time step, and we assume that there are sufficiently large episode K for recipient agents to learn the perceived transition kernels and reward functions. Notice that in this case, the perceived transition kernels and reward functions are consistent across time steps, since the same permutation attack Σ_i is applied at every time step.

Learning: The perceived transition kernel (same in every time step) is $P^{d,i} = \Sigma_i P^i \Sigma_i^T$, where P^i is the ground-truth transition kernel of agent i. The perceived reward function (same in every time step) is $R_i^d = \Sigma_i R_i$, where R_i is the ground-truth reward function of agent i.

C. Dynamic programming

In both cases of training-time camouflage attacks, after every agent i learns the series of perceived transition kernels $\{P_{t \to t+1}^{d,i}\}_{t=0}^{T-1}$ and $\{R_{i,t}^d\}_{t=1}^{T}$, we use the dynamic programming to find the delusional "optimal" policy $\pi_i^{d,*} = \{\pi_{i,t}^{d,*}(s)\}_{t=0}^{T-1}, s \in \mathcal{S}_i$. Notice that $\pi_i^{d,*}$ is the "optimal policy" in the eyes of any recipient agent i since recipient agents are unaware of attacks during their learning.

Algorithm 1 Dynamic programming under camouflage attacks

```
Input: $\mathcal{S}_i,\ \mathcal{A}_i,\ T,\ \{P_{t\to t+1}^{d,i}\}_{t=0}^{T-1}, \{R_{i,t}^d\}_{t=1}^{T}$
Initialize $V_{i,T}^{d,*} = 0$
for $t = T-1, T-2, \ldots, 0$ do
    for $s_i \in \mathcal{S}_i$ do
        $V_{i,t}^{d,*}(s) \leftarrow \max_{a_i \in \mathcal{A}_i} R_{i,t+1}^d(s_i, a_i) + \sum_{s_i'} P_{t\to t+1}^{d,i}(s_i'|s_i, a_i) V_{i,t+1}^{d,*}(s_i')$
        $\pi_{i,t}^{d,*}(s) \leftarrow \arg\max_{a_i \in \mathcal{A}_i} R_{i,t+1}^d + \sum_{s_i'} P_{t\to t+1}^{d,i}(s_i'|s_i, a) V_{i,t+1}^{d,*}(s_i')$
    end
end
Output $\{\pi_{i,t}^{d,*}(s_i)\}_{t=0}^{T-1}$
```

V. Performance analysis of camouflage attacks

A camouflage attack is arguably a more practical form of adversarial attack since it only requires the attackers to change the appearances of the objects the attackers directly control. So different victims will have correlated or the same observations of these camouflaged objects. In contrast, the optimal state perception attacks would require the attackers to change the observations of different victims to possibly different delusions. The analytical results on both the test-time cases and training-time cases bound the gaps between expected rewards earned under the camouflage attacks and under the state perception attacks.

In this section, we assume that different victims have the same observations of the camouflaged objects and we do not impose cost constraints on the attacks. We start with a lemma about imposing an equality constraint on the optimization variables.

Lemma V.1. *Consider n functions $\{f_i\}_{i=1}^n$, where $i = 1, 2,\ \ldots,\ n$, and the following two optimization problems:*

$$\min_{x_1, x_2, \ldots, x_n} \sum_{i=1}^{n} f_i(x_i) \tag{6}$$
$$\textit{subject to } x_1 = x_2 = \cdots = x_n;$$

and

$$\min_{x_1, x_2, \ldots, x_n} \sum_{i=1}^{n} f_i(x_i). \tag{7}$$

Let x^{**} be the optimal solution of (6) and o_1 be the optimal objective value of (6). Let $(x_1^*, x_2^*, \ldots, x_n^*)$ be the optimal solution of (7) and o_2 be the optimal objective value of (7). Assume that there exist constants C_j's, $j = 1, 2,\ \ldots,\ n$, such that for every j,

$$\sum_{i=1, i\neq j}^{n} (f_i(x_j^*) - f_i(x_i^*)) \leq C_j.$$

Then we have $o_2 \leq o_1 \leq o_2 + \min_j \{C_j\}$.

Proof. : Because (6) has one additional constraint, $o_2 \leq o_1$. Given an arbitrary index j, $j = 1 \ldots n$, we have:

$$o_1 = \sum_{i=1}^{n} f_i(x^{**}) \leq \sum_{i=1}^{n} f_i(x_j^*).$$

So for every j, we have

$$\begin{aligned} o_1 - o_2 &\leq \sum_{i=1}^{n} f_i(x_j^*) - \sum_{i=1}^{n} f_i(x_i^*) \\ &= \sum_{i=1, i\neq j}^{n} (f_i(x_j^*) - f_i(x_i^*)) \leq C_j. \end{aligned}$$

Therefore $o_2 \leq o_1 \leq o_2 + \min_j \{C_j\}$. □

We use Lemma $V.1$ to prove the bounded gap between the global rewards gained under test-time state perception attacks and under test-time camouflage attacks.

A. Performance of test-time camouflage attacks

Theorem V.2 (Test-time comparison)**.** *Consider m attackers and n recipient agents, for one single time step t (from time index $t-1$ to time index t). Assume the recipients share the same state space $\mathcal{S}$, the same action space $\mathcal{A}$, the same* *probability transition matrices $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$, and the same reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. Let the observation functions h_i be identical for every agent i, so that the camouflaged observations are the same for every recipient agent i, i.e. $s_{d,t-0.5,i}$'s are equal. We assume that the optimal policy for each recipient agent is the same and the recipients work independently from each other. We use $\pi_t^* : \mathcal{S} \rightarrow \mathcal{A}$ to denote the shared optimal policy of a recipient agent at time step t. Within time step t, let the expected reward gained by individual agent i be ER. Let the total rewards of all recipients gained under the optimal camouflaged attack be TR_t^{ca} and the total reward gained under the optimal state-perception attack be TR_t^{spa}.*

Assume that for every pair of two different recipient agents (i, j), $i, j = 1 \dots n$, for every pair of actual states $(s_{a,t-1,i}, s_{a,t-1,j})$ of recipient i and recipient j, the rewards gained for agent i under delusional state perceptions at time step t satisfy

$$ER(s_{a,t-1,i}, \pi_t^*(s^*_{d,t-0.5,j})) - ER(s_{a,t-1,i}, \pi_t^*(s^*_{d,t-0.5,i})) \leq C_{ij},$$

for some small constant C_{ij}, where $s^*_{d,t-0.5,j}$ and $s^*_{d,t-0.5,j}$ are the most-damaging delusional state perceptions that minimize the reward for agent j and i respectively, and $ER(s_{a,t-1,i}, \pi_t^*(s_{d,t-0.5,\cdot}))$ is the expected reward recipient agent i will get using the policy corresponding to a delusional state perception $(s_{d,t-0.5,\cdot})$. Then

$$TR_t^{spa} \leq TR_t^{ca} \leq TR_t^{spa} + \min_j \sum_{i=1, i \neq j}^{n} \{C_{ij}\}.$$

Proof. : We use Lemma V.1 to prove Theorem V.2. For each recipient agent i, $1 \leq i \leq n$, we let function f_i be the expected reward recipient agent i gets under its true states $s_{a,t-1,i}$ and agent i's delusional observation $s_{d,t-0.5,i}$.

In this setting, the variable x_i in the optimization problems (6) and (7) is the delusional observation $s_{d,t-0.5,i}$ of the i-th recipient. In (6) which corresponds to the camouflage attack, $f_i(x_i) = ER(s_{a,t-1,i}, \pi_t^*(s_{d,t-0.5,i}))$ and we require $s_{d,t-0.5,i}$ to be equal across different agents i's. In (7) which corresponds to a "free" state perception attack, $f_i(x_i) = ER(s_{a,t-1,i}, \pi_t^*(s_{d,t-0.5,i}))$, but we will not require $s_{d,t-0.5,i}$ to be the same across different agents i's. Because

$$ER((s_{a,t-1,i}, \pi_t^*(s^*_{d,t-0.5,j})) - ER(s_{a,t-1,i}, \pi_t^*(s^*_{d,t-0.5,i})) \leq C_{ij},$$

by applying Lemma V.1, we have $TR_t^{spa} \leq TR_t^{ca} \leq TR_t^{spa} + \min_j \sum_{i=1, i \neq j}^{n} \{C_{ij}\}$. □

B. Performance of the training-time one-time camouflage attack

To compare the effect of training-time state perception attacks and training-time camouflaged attacks, we assume that agents share the same state space such that $\mathcal{S}_i = \mathcal{S}, \forall \mathcal{S}_i$ with $|\mathcal{S}| = S$, the same action space $\mathcal{A}_i = \mathcal{A}, \forall \mathcal{A}_i$ with $|\mathcal{A}| = A$, but each agent i has its own transition kernel P_i and reward function r_i. For every agent i, let the ground-truth infinite MDP environment be $E = (\mathcal{S}, \mathcal{A}, P_i, r_i, T, \gamma)$, and the perceived infinite MDP environment be $E^d = (\mathcal{S}, \mathcal{A}, P_i^d, R_i^d, T, \gamma)$, wherein $T \rightarrow \infty$, and $\gamma \in [0, 1]$. γ is the discount factor. In the version of finite MDP, we assume T is finite and $\gamma = 1$.

Let $\pi_{d,i}(s|\dagger), \forall s \in \mathcal{S}$, denote the delusional "optimal" policy for agent i. Every recipient agent i learns $\pi_{d,i}(\cdot|\dagger)$ in the perceived environment E^d during training under the attack strategy †. Here $\cdot$ is a dummy variable which means any state s. We keep the notation $\cdot$ to make our citations of policies consistent with traditional notations. We compare the **tested** V-values $V_{i,T}^{\pi_{d,i}(\cdot|\dagger)} \in \mathbb{R}^S$ by following the delusional "optimal" policy in the ground-truth MDP envrionment E, in which

$$V_i^{\pi_{d,i}(\cdot|\dagger)}(s) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_i(s_k, a_k^\dagger) | s_0 = s, a_k^\dagger \sim \pi_{d,i}(s_k|\dagger)].$$

In the finite-horizon version,

$$V_{i,T}^{\pi_{d,i}(\cdot|\dagger)}(s) = \mathbb{E}[\sum_{k=0}^{T} r_i(s_k, a_k^\dagger) | s_0 = s, a_k^\dagger \sim \pi_{d,i}(s_k|\dagger)].$$

The following theorem compares the gap between tested V-values guided by delusional polices trained under the one-time state perception attack and under the one-time camouflage attack. We assume E and E^d finite MDPs in Theorem V.3.

Theorem V.3 (Comparisons under one-time camouflage attack)**.** *Consider m attackers and n recipient agents present in a finite MDP with horizon T. In the one-time state perception attacks, the set of attack strategies $\{\dagger\}$ is the set of state perception probability matrices $\{\bar{X}_i\} = \{[x_{s_{a,i,t^*} \rightarrow s_{d,i,t^*}}]_{s_{a,i,t^*}, s_{d,i,t^*} \in \mathcal{S}}\}$, in which $\{x_{s_{a,i,t^*} \rightarrow s_{d,i,t^*}}\}$ are subject to the conditions in (2).*

Assume that for every pair of two different recipient agents (i, j), for every state s, the tested V-values gained for agent i with its policy trained under the training-time state-perception-attack strategy $\bar{X}_i^*$ satisfy

$$V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_j^*)}(s) - V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_i^*)}(s) \leq C_{ij},$$

for some small constant C_{ij}, where $\bar{X}_j^*$ and $\bar{X}_i^*$ are the most damaging state perception probability matrices that minimize the tested V-values for agent j and i respectively. For one-time camouflage attacks, we require the optimal camouflage attack strategy $\bar{X}^*$ to be the same across all recipient agents. Then,

$$\sum_{i=1}^{n} V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_i^*)} \leq \sum_{i=1}^{n} V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}^*)} \leq \sum_{i=1}^{n} V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_i^*)} + \min_j \sum_{i=1, i \neq j}^{n} \{C_{ij}\}$$

Proof. : We use Lemma V.1 to prove V.3. For each recipient agent i, we let the function $\{f_i\}_{i=1}^n$ be the tested V-value recipient agent i gets at time index T, under the attack strategy $\bar{X}_i$. The variable x_i in the optimization problems (7) and (6) are the state perception probability matrices $\bar{X}_i$. In (6) which corresponds to the camouflage attack, $f_i(x_i) = V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_i)}$ and we require $\bar{X}_i$ to be equal across different agents i's. In (7), which corresponds to a state perception attack, $f_i(x_i) = V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_i)}$, but we don't require $\bar{X}_i$ to be the same across all agents i's. Since $V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_j^*)} - V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_i^*)} \leq C_{ij}$, by applying Lemma V.1, we have the direct proof of Theorem V.3:

$$\sum_{i=1}^{n} V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_i^*)} \leq \sum_{i=1}^{n} V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}^*)} \leq \sum_{i=1}^{n} V_{i,T}^{\pi_{d,i}(\cdot|\bar{X}_i^*)} + \min_j \sum_{i=1, i \neq j}^{n} \{C_{ij}\}.$$

□

C. Performance of the training-time permutation attacks

We further prove analytical results of training-time permutation camouflage attacks with infinite MDPs. We start by proving two lemmas that the Bellman optimality operator $\mathbb{T}$ is a contraction and that the perceived V-value is a permutation of the ground-truth V-value. Recall that we use $V_{i,t}^{d,*}$ to denote the delusional optimal V-value of agent i at time index t during the training under attacks. Since in this case we are assuming an infinite MDP, we omit the time index t and use $V_i^{d,*}$ to denote the delusional optimal V-value at the end of the training which was under attacks. The same case applies to the notation of the ground-truth optimal V-values V_i^*.

Then, we use those two lemmas to prove the convergence of value iteration in the perceived environment E^d under permutation camouflage attacks.

Lemma V.4. *(Contraction of Bellman optimality operator) Let $\mathbb{T}$ be the Bellman optimality operator. Given any state $s \in \mathcal{S}$,*

$$\mathbb{T}(V_i)(s) = \max_{a\in\mathcal{A}} \ \bar{r}_i(s,a) + \gamma \sum_{s'\in\mathcal{S}} P_i(s'|s,a)V_i(s'),$$

where $\bar{r}_i$ is the expected value of r_i. The optimality operator $\mathbb{T}$ is a contraction in the l_∞-norm with the contraction factor γ, i.e.,

$$\|\mathbb{T}(V_i) - \mathbb{T}(V_i')\|_\infty \leq \gamma \|V_i - V_i'\|_\infty.$$

Proof. Let π denote the greedy policy with respect to the Bellman optimality operator $\mathbb{T}(V_i)$, such that π satisfies:

$$\mathbb{T}(V_i)(s) = \mathbb{T}^\pi(V_i)(s) = \bar{r}_i(s,\pi(s)) + \gamma \sum_{s'\in\mathcal{S}} P_i(s'|s,\pi(s))V_i(s'),$$

where $\pi(s) = \arg\max_a(\bar{r}_i(s,a) + \gamma \sum_{s'\in\mathcal{S}} P_i(s'|s,a)V_i(s'))$. Let $V_i, V_i' \in \mathbb{R}^S$. Given a state $s \in \mathcal{S}$, we have that:

$$\begin{aligned}\|\mathbb{T}(V_i)(s) - \mathbb{T}(V_i')(s)\|_\infty &= \|\bar{r}_i(s,\pi(s)) + \gamma \sum_{s'\in\mathcal{S}} P_i(s'|s,\pi(s))V_i(s') \\ &\quad - \Big(\bar{r}_i(s,\pi(s)) + \gamma \sum_{s'\in\mathcal{S}} P_i(s'|s,\pi(s))V_i'(s')\Big)\|_\infty \\ &= \gamma \|\sum_{s'\in\mathcal{S}} P_i(s'|s,\pi(s))(V_i(s') - V_i'(s'))\|_\infty \\ &\leq \gamma \sum_{s'\in\mathcal{S}} P(s'|s,\pi(s))\|(V_i(s') - V_i'(s'))\|_\infty \\ &= \gamma \|V_i - V_i'\|_\infty.\end{aligned}$$

□

Lemma V.5. *[Permutation of value functions] Consider m attackers and n agents present in an infinite MDP. Attackers perform permutation attacks with the same attack strategy Σ_i, where Σ_i is a permutation matrix with S rows and columns, at every time step t on the individual agent i. Using value iteration, the recipient agent i perceives its optimal V-value $V_i^{d,*}$ as a **permutation** of the ground-truth optimal V-values V_i^*, i.e., $V_i^{d,*} = \Sigma_i V_i^*$*

Proof. We prove Lemma V.5 by induction. Let $V_{i,0}^* = 0$ and $V_{i,0}^{d,*} = \Sigma_i V_{i,0}^*$ be the initial approximations. Using value iteration to obtain both the true V-values and the perceived V-values, we have the next true approximation to be $V_{i,1}^* = \max_a(r_i + \gamma P_i V_{i,0}^*)$, and next perceived approximation $V_{i,1}^{d,*}$ to be:

$$\begin{aligned}V_{i,1}^{d,*} &= \max_a(\Sigma_i r_i + \gamma \Sigma_i P_i \Sigma_i^T V_{i,0}^{d,*}) \\ &= \Sigma_i(\max_a(r_i + \gamma P_i \Sigma_i^T \Sigma_i V_{i,0}^*)) \\ &= \Sigma_i(\max_a(r_i + \gamma P_i V_{i,0}^*)) = \Sigma_i V_{i,1}^*\end{aligned}$$

For the induction step, assume that at time step k, $V_{i,k}^{d,*} = \Sigma_i V_{i,k}^*$. Then, the next true approximation is $V_{i,k+1}^* = \max_a(r_i + \gamma P_i V_{i,k}^*)$, and the next perceived approximation $V_{i,k+1}^{d,*}$ is:

$$\begin{aligned}V_{i,k+1}^{d,*} &= \max_a(\Sigma_i r_i + \gamma \Sigma_i P_i \Sigma_i^T V_{i,k}^{d,*}) \\ &= \Sigma_i(\max_a(r_i + \gamma P_i \Sigma_i^T \Sigma_i V_{i,k}^*)) \\ &= \Sigma_i(\max_a(r_i + \gamma P_i V_{i,k}^*)) = \Sigma_i V_{i,k+1}^*\end{aligned}$$

□

Theorem V.6 (Convergence analysis under permutation camouflage attacks)**.** *For every agent i, let V_i^* be the optimal V-value of the ground-truth MDP E, and let $V_i^{d,*}$ be the optimal V-value of the perceived MDP E^d, which were both found by value iterations. The algorithm of value iteration on E^d is given in Algorithm 2.*

Assuming the iteration number is n, the gap between the true optimal value V_i^* and the n-th estimation of the perceived V-value $V_{i,n}^d$ in l_∞-norm is bounded by

$$\|V_i^* - V_{i,n}^d\|_\infty \leq \frac{\gamma^n}{1-\gamma}\|V_{i,1}^d - V_{i,0}^d\|_\infty + \|V_i^* - V_i^{d,*}\|,$$

and

$$\|V_i^* - V_{i,n}^d\|_\infty \geq \|V_i^* - V_i^{d,*}\| - \frac{\gamma^n}{1-\gamma}\|V_{i,1}^d - V_{i,0}^d\|_\infty, \quad (8)$$

wherein $V_{i,0}^d$ is the initial approximation and $V_{i,1}^d$ is the approximate perceived V-value by applying $\mathbb{T}$ once.

Proof. By the triangle inequality,

$$\|V_i^* - V_{i,n}^d\|_\infty \leq \|V_i^* - V_i^{d,*}\|_\infty + \|V_i^{d,*} - V_{i,n}^d\|_\infty$$

To bound $\|V_i^{d,*} - V_{i,n}^d\|_\infty$, we make a convergence analysis on the value iteration. Let $\mathbb{T}$ denote the Bellman optimality operator and apply $\mathbb{T}$ on E^d. Since $\mathbb{T}$ is a contraction map, $V_i^{d,*}$ is the fixed point of $\mathbb{T}$, i.e., $\mathbb{T}(V_i^{d,*}) = V_i^{d,*}$. Notice that with an arbitrary step m, by the contraction property of $\mathbb{T}$, we have that:

$$\begin{aligned}\|V_{i,m+1}^d - V_{i,m}^d\|_\infty &= \|\mathbb{T}(V_{i,m}^d) - \mathbb{T}(V_{i,m-1}^d)\|_\infty \\ &\leq \gamma \|V_{i,m}^d - V_{i,m-1}^d\|_\infty \\ &\leq \cdots \\ &\leq \gamma^m \|V_{i,1}^d - V_{i,0}^d\|_\infty,\end{aligned}$$

and therefore, we can bound the second part of the triangle inequality $\|V_i^{d,*} - V_{i,n}^d\|_\infty$ by creating a telescoping series:

$$\begin{aligned}\|V_i^{d,*} - V_{i,n}^d\|_\infty &\leq \|V_{i,n+1}^d - V_{i,n}^d\|_\infty + \|V_{i,n+2}^d - V_{i,n+1}^d\|_\infty \\ &+ \dots + \|V_{i,n+l}^d - V_{i,n+l-1}^d\|_\infty + \|V_i^{d,*} - V_{i,n+l}^d\|_\infty \\ &\leq (\gamma^n + \gamma^{n+1} + \dots + \gamma^{n+l-1})\|V_{i,1}^d - V_{i,0}^d\|_\infty \\ &+ \gamma^{n+l}\|V_i^{d,*} - V_{i,0}^d\|_\infty \\ &\leq \frac{\gamma^n}{1-\gamma}\|V_{i,1}^d - V_{i,0}^d\| + \gamma^{n+l}\|V_i^{d,*} - V_{i,0}^d\|_\infty\end{aligned}$$

Let $l \to \infty$, we have that

$$\|V_{i,n}^d - V_i^{d,*}\| \leq \frac{\gamma^n}{1-\gamma}\|V_{i,1}^d - V_{i,0}^d\|.$$

We use Lemma V.5 to bound $\|V_i^* - V_i^{d,*}\|$, and $V_i^{d,*} = \Sigma_i V_i^*$. Similarly, by the triangle equality

$$\|V_i^* - V_{i,n}^d\|_\infty \geq \|V_i^* - V_i^{d,*}\|_\infty - \|V_i^{d,*} - V_{i,n}^d\|_\infty,$$

we further have the proof of (8).

□

Algorithm 2 Value iteration under permutation attacks

```
Input: Infinite camouflaged horizon MDP($\mathcal{S}, \mathcal{A}, P_i^d, R_i^d, \gamma$), iteration number: $n$
for $k = 0, 1, \dots, n-1$ do
    for $s \in \mathcal{S}$ do
        $V_{i,k+1}^{d,*} \leftarrow \mathbb{T}(V_{i,k}^{d,*})$
        $\pi_{i,k}^{d,*} \leftarrow$ greedy policy of $V_{i,k}^{d,*}$
    end
end
Output: $V_{i,n}^{d,*}$, $\pi_{i,n}^{d,*}$
```

Our last theorem is to bound the gap between tested V-values guided by delusional policies obtained under state-perception permutation attacks and under camouflage permutation attacks. We assume E and E^d infinite MDPs in Theorem V.7.

Theorem V.7 (Comparisons of training-time permutation camouflage attacks)**.** *Consider m attackers and n recipient agents present in an infinite MDP. In the state perception permutation attacks, the set of attack strategies $\{\dagger\}$ is the set of state perception permutation matrices $\{\Sigma_i\}$, which is a set of permutation matrices with S rows and columns.*

Assume that for every pair of two different recipient agents (i, j), for every state s, the tested V-values gained by agent i with its policy trained under state-perception-attack strategy Σ_i^* satisfy

$$V_{i,T}^{\pi_{d,i}(\cdot|\Sigma_j^*)} - V_{i,T}^{\pi_{d,i}(\cdot|\Sigma_i^*)} \leq C_{ij},$$

for some small constant C_{ij}, where Σ_j^* and Σ_i^* are the most damaging permutation matrices that minimize the tested V-value for agent j and i respectively. In permutation camouflage attacks, we require the optimal permutation attack strategy Σ^* to be the same across all recipient agents. Then,

$$\sum_{i=1}^n V_{i,T}^{\pi_{d,i}(\cdot|\Sigma_i^*)} \leq \sum_{i=1}^n V_{i,T}^{\pi_{d,i}(\cdot|\Sigma^*)} \leq \sum_{i=1}^n V_{i,T}^{\pi_{d,i}(\cdot|\Sigma_i^*)} + \min_j \sum_{i=1, i\neq j}^n \{C_{ij}\}$$

Proof. We prove by Lemma V.1. In the training, the camouflage permutation attacks permit only one minimizer, Σ. However, for state perception permutation attacks, there are n minimizers allowed: (Σ_1, Σ_2,…,Σ_n). □

VI. Numerical Results

We perform numerical results on both test-time and training-time camouflage attacks under various game settings with finite time T. In the test-time case, we compare the total rewards of all recipients without attack, under the camouflage attack, and under the state perception attack. Results indicate that recipients gain significantly smaller rewards under the camouflage attacks compared to the case with no attacks. The reward gained under the state perception attacks is smaller, but not significantly, than the more practical camouflage attack. For cost-constrained camouflage attacks, as the attack budget increases, the gained reward becomes smaller. Our framework works for general m-attacker-n-recipient scenarios. In the training-time case, we perform a one-time camouflage attack during the learning of the multiple-agent system and observe a significant drop in reward resulting from the camouflage attack.

A. Camouflage orientations (Test-time)

In the first experiment, there are 2 recipients and 2 attackers playing in the MG. The recipients share the same state space $\mathcal{S}$, which is a ring containing 3 different states: 0, 1, and 2. They also share the same action space $\mathcal{A}$, the probability transition P, and the same reward function R. The action space $\mathcal{A}$ is composed of three actions: go left, go right, and stay. For actions **left** and **right**, the recipient agent has a 0.8 probability of moving in the intended direction and a 0.2 probability of moving in the opposite direction. For **stay**, the recipient agent has a 0.8 probability of staying at the current state and a 0.1 possibility of moving to the right or left. The reward function in the experiment depends on the current state, and the next state: $R(s_{t-1}, s_t)$, which assigns a fixed positive reward to the recipient agents, which is displayed in the table below:

$t \downarrow t+1 \rightarrow$	s_0	s_1	s_2
s_0	3.0	10.6	1.0
s_1	10.0	1.0	0.0
s_2	1.0	0.0	11.6

The attacks camouflage the orientation of the ring by rotating it counter-clockwisely for respectively 1 step, 2 steps, and 3 steps. The camouflaged orientation after a 3-step rotation is the same as the true orientation. For every attack, recipients' perceptions of their real positions are based on the camouflaged ring, as described in Figure 2. In Figure 3, we compared the expected global rewards of recipients for time index from 0 to 5, under camouflage attacks, state perception attacks, and without attacks. In this figure, the x-axis is the time index, and the y-axis is the total expected rewards the recipient agents gained from $t = 0$ to the current time index. With camouflage attacks, the reward gained is 34.4% of that achieved without attack. With state perception attacks (where attackers can freely fool each recipient into desired delusional states), the reward is about 33.1% of that without attack.

B. Camouflage attackers' real positions (Test-time)

In this case, the recipients and the attackers move on a square $q \times q$ chessboard. The position of either a recipient or

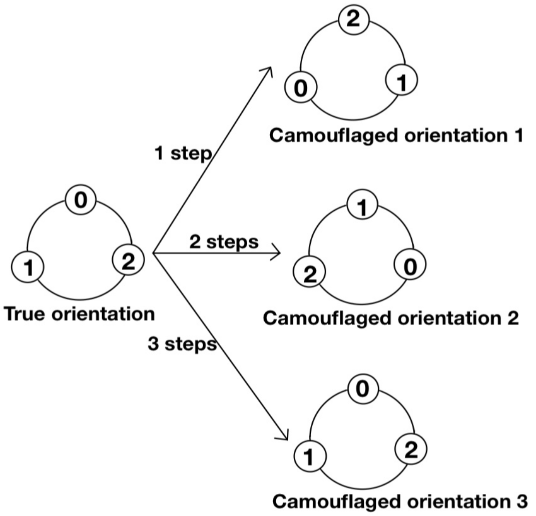
2
0
1
Camouflaged orientation 1
1 step
0
1
2
True orientation
2 steps
1
2
0
Camouflaged orientation 2
3 steps
0
1
2
Camouflaged orientation 3

Figure 2: Illustration: camouflage attacks on a ring.

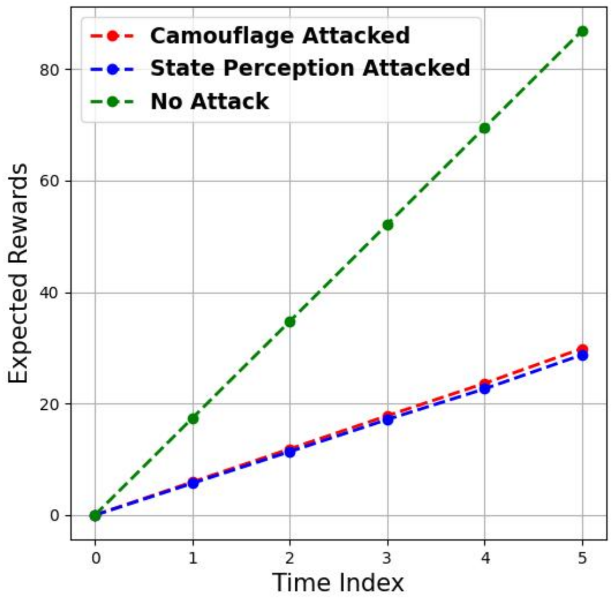

Figure 3: Ring topology. Comparison of expected global rewards between free state perception attacks and camouflage attacks.

an attacker can be denoted as (i, j), where i $(0 \leq i \leq q-1)$ records the row index, and j $(0 \leq j \leq q-1)$ records the column index on the chessboard. We made 2 experiments with a square chessboard: the first one has $q = 3$, and the second one has $q = 2$. For each attacker, its position is fixed and it can only attack if any recipient moves to its location. Neither attackers nor recipients can move beyond the boundaries of the chessboard. The recipients have the same state space $\mathcal{S}$, the same action space $\mathcal{A}$, and the same reward function R. Recipients can move up, down, right, or left. The attackers' positions cannot move during the game.

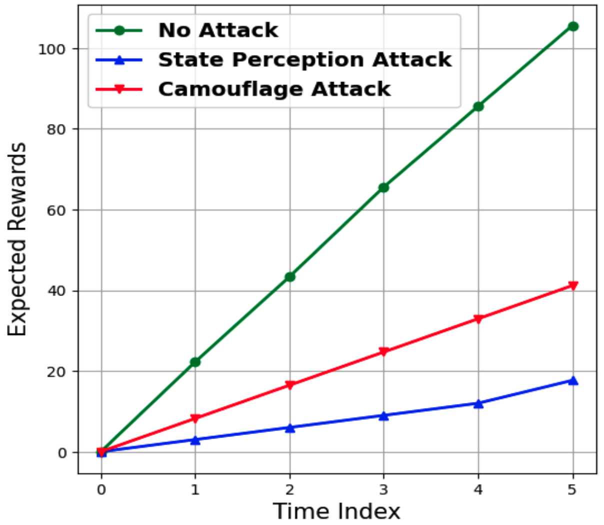

Figure 4: 3×3 chessboard. Comparison between state perception attack and camouflage attack, with fixed attackers at (1,1) and (2,1), 3 recipients and 2 attackers.

For simplicity, the probability of recipients moving along the indicated direction of an action $a \in \mathcal{A}$ is set to be 1. The reward $R(s_{t-1}, s_t)$ for entering all possible chessboard positions is set to be 5.0 except for $(0, 1)$, whose reward is 10.0. If any recipient enters any of the attacker's positions, the reward is set to 1.0. For example, with $q = 3$, if the fixed positions of the attackers are at $(1, 1)$ and $(2, 1)$, then the reward function can be displayed in the following table. The i, j here means the indices of the square the recipient is entering:

$j \downarrow i \rightarrow$	0	1	2
0	5.0	10.0	5.0
1	5.0	1.0	5.0
2	5.0	1.0	5.0

In the first experiment, 3 recipients and 2 attackers played on a 3×3 chessboard, and in the second experiment, 2 recipients and 1 attacker played on a 2×2 chessboard. In Figures 4 and 5, we compared the expected global rewards of recipients for time index from 0 to 5, under camouflage attacks, state perception attacks, and without attacks.

In Figure 4, with the fixed attackers at (1,1) and (2,1) on the chessboard, the global reward gain achieved under camouflage attacks is 39.0% of the reward achieved without attack. The reward gain under state perception attacks is roughly 16.7%. In Figure 5, the expected global gained reward overall possible real attackers' positions after 5 camouflage attacks is 47.3% of the case without attack, and the expected gained reward after 5 state perception attacks is 43.6% of the case without attack.

C. *Cost constrained camouflage attacks (Test-time)*

With 3 recipients and 2 attackers in the same 3×3 chessboard setting as VI-B, we add the cost constraint to attackers at every time step when they perform a camouflage attack. The cost of every attacker is the distance between its real position and the target camouflage position. We define the distance between two positions as the sum of their row and

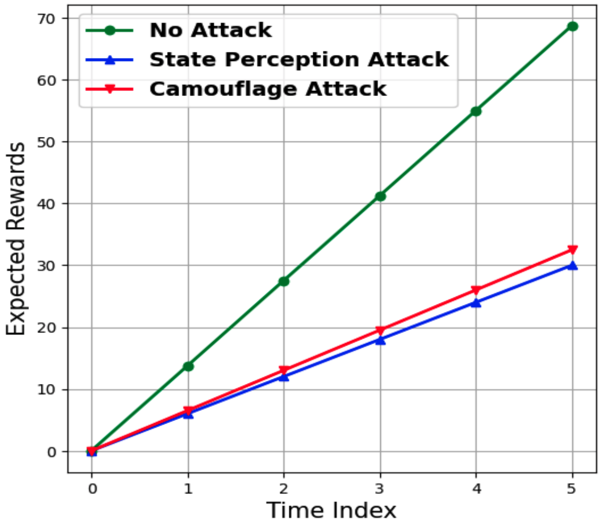

Figure 5: 2×2 chessboard. Comparison between state perception attack and camouflage attack, 2 recipients and 1 attacker.

column index absolute differences. Within every time step t, the shared budget is refilled to a fixed budget. We choose the following sequence of fixed budgets $\{1, 2, 3, 4, 6, 12\}$ for tests. In Figure 6, we compare the expected global reward gain under different budgets. It turns out that the higher the budget, the fewer the reward gains. When the budget reaches 6, the performance of the cost constrained camouflage attack is the same as the optimal camouflage attack.

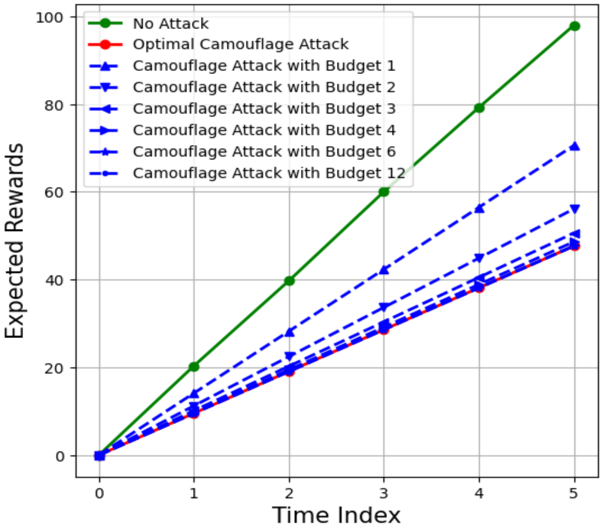

Figure 6: Comparison between camouflage attacks with different budgets.

D. *Camouflage orientations (Training-time)*

In the training-time camouflage attack experiment, 2 recipients play in the MG. The recipients share the state space $\mathcal{S} = \{(s_1, s_2)\}$, where $s_1, s_2 \in \{0, 1, 2\}$ on a circle. The joint action space $\mathcal{A}$ contains 9 joint actions {(left,left), (left,right), (left,stay), (right,left), (right, right), (right,stay), (stay,left), (stay,right), (stay,stay)}. The ground-truth transition probabilities of every action {left, right, stay} are displayed in the following tables. The transition probabilities of joint actions, $P_{(a_1,a_2)}$ are the tensor products of P_{a_1} and P_{a_2}, for example, $P_{L,L} = P_L \otimes P_L$. The random reward function of each agent is $r_i(s, a)$, depends on the state action pair (s, a) and is distributed from the normal distribution $r_i \sim \mathcal{N}(\mu_i, \sigma_i)$. μ_i and σ_i are chosen by Table III.

0	0.25	0.75
0.85	0	0.15
0.25	0.75	0

0	0.85	0.15
0.15	0	0.85
0.85	0.15	0

0.9	0.05	0.05
0.05	0.9	0.05
0.05	0.05	0.9

Table I: $P_{left},\ P_{right},\ P_{stay}$

$a \downarrow s_t \rightarrow$	0	1	2
left	10, .02	2, .01	20, .01
right	5, .03	20, .02	4, .01
stay	1, .01	10, .03	40, .02

Table II: (μ_1, σ_1) for individual reward r_1

$a \downarrow s_t \rightarrow$	0	1	2
left	4, .01	20, .01	40, .02
right	20, .02	2, .01	10, .03
stay	5, .03	10, .02	1, .01

Table III: (μ_2, σ_2) for individual reward r_2

The reward function of the MG is the total reward of two agents: $r_1 + r_2$.

We let $T = 7$. At $t = 3$, the attacker performs the one-time camouflage attack to confuse agents' observations on the circle orientation. There is a fixed probability distribution $p(s_d|s_a)$ that agents would perceive their true states as false states. The elements of the camouflage probability matrix $\bar{X}$ were chosen randomly between 0 and 1, but the matrix has both row-sums and column-sums to be 1. One example of camouflage probability matrix is shown in Table IV. The camouflage probability of the 2-agent system is $\bar{X} \otimes \bar{X}$.

$s_a \rightarrow s_d \downarrow$	**0**	**1**	**2**
0	0	0.68578239	0.31421761
1	0.31421761	0	0.68578239
2	0.68578239	0.31421761	0

Table IV: Camouflage probability

Under the one-time camouflage attack at $t = 3$, the agents learn the camouflaged transition probabilities $P^d_{2\to3}$, $P^d_{3\to4}$, and the camouflaged reward function R^d_3. At other time indices, agents learn the ground-truth transition probabilities and reward functions.

We compare the sum of expected reward across all agents by following the attacked optimal policies $\pi_i^{d,*}, i = 1, 2$ and the true optimal policies $\pi_i^*, i = 1, 2$ in Figure 7. Compared with the ground-truth V-values, the camouflaged V-values start to drop at $t = 4$ since $\pi_{i,3}^{d,*} \neq \pi_{i,3}^*$ due to the camouflage attack at $t = 3$. The drop of the expected reward of the state $(0, 0)$, for example, is about 28.5% of the ground-truth $V^*(0, 0)$.

We also did an experiment on a single-agent system, with $\mathcal{S} = \{0, 1, 2\}$, $\mathcal{A} = \{left,\ right,\ stay\}$, and the same P_a and r as described in the 2-agent system. The numerical result proves to be similar in Figure 8.

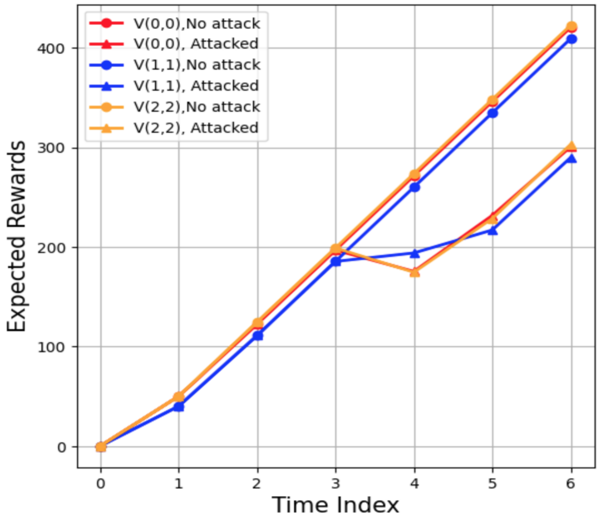

Figure 7: Expected reward under a training-time camouflage attack-multiple agents

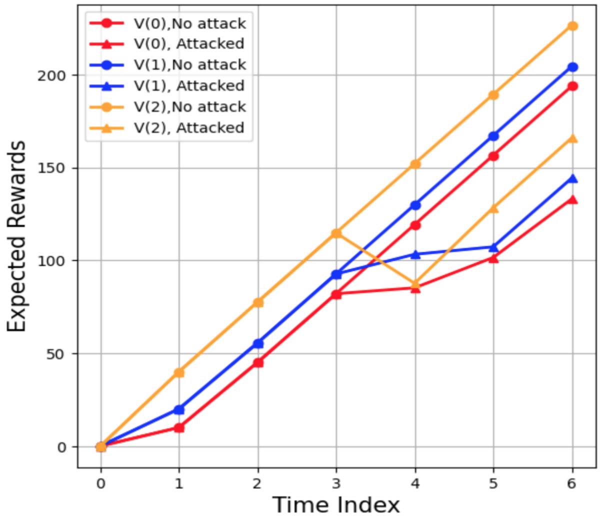

Figure 8: Expected reward under a training-time attack-single-agent

E. Camouflage attack on Nash equilibrium policy (Test-time)

In the last experiment, we would like to explore the effect of camouflage attacks on Nash equilibrium solutions. We assume a two-agent general sum Markov game, with shared state space $\mathcal{S} = \{0, 1\}$. The action spaces of Agent 1 and Agent 2 both consist of two elements $\mathcal{A}_1 = \mathcal{A}_2 = \{0, 1\}$. At $s = 0$, the shared state transits to $s = 1$ with probability $1/4$ only if agents take the joint action $(1, 1)$. Otherwise, the state of the game stays at $s = 0$. At $s = 1$, the shared state stays the same no matter what join action agents take.

The reward functions (r_1, r_2) both depend on s and the joint action $\boldsymbol{a} = (a_1, a_2)$, which are displayed as the following

$a_1 \downarrow a_2 \rightarrow$	0	1
0	$(-1, 0)$	$(-1, 0)$
1	$(1, -1)$	$\begin{cases}(-20, 3) \text{ with prob } 1/4 \\ (0, 0) \text{ with prob } 3/4\end{cases}$

Table V: (r_1, r_2) of $s = 0$

$a_1 \downarrow a_2 \rightarrow$	0	1
0	$(2, 1)$	$(2, 0)$
1	$(2, -1)$	$\begin{cases}(-20, 3) \text{ with prob } 1/4 \\ (0, 0) \text{ with prob } 3/4\end{cases}$

Table VI: (r_1, r_2) of $s = 1$

By initializing the V-values $V^{\boldsymbol{\pi}}_{i,T}(s) = 0$ for any agent i, of any state s, and any joint policy $\boldsymbol{\pi}$, we can obtain the Q-values of Agent 1 at time index $T - 1$ as:

$$\begin{aligned}
Q^{\boldsymbol{\pi}}_{1,T-1}(0, (0, 0)) &= -1 + V^{\boldsymbol{\pi}}_{1,T}(0), \\
Q^{\boldsymbol{\pi}}_{1,T-1}(0, (0, 1)) &= -1 + V^{\boldsymbol{\pi}}_{1,T}(0), \\
Q^{\boldsymbol{\pi}}_{1,T-1}(0, (1, 0)) &= 1 + V^{\boldsymbol{\pi}}_{1,T}(0), \\
Q^{\boldsymbol{\pi}}_{1,T-1}(0, (1, 1)) &= -5 + 1/4V^{\boldsymbol{\pi}}_{1,T}(1) + 3/4V^{\boldsymbol{\pi}}_{1,T}(0).
\end{aligned}$$

Similarly, for Agent 2, we have Q-values at $T - 1$ as:

$$\begin{aligned}
Q^{\boldsymbol{\pi}}_{2,T-1}(0, (0, 0)) &= 1 + V^{\boldsymbol{\pi}}_{2,T}(0), \\
Q^{\boldsymbol{\pi}}_{2,T-1}(0, (0, 1)) &= -1 + V^{\boldsymbol{\pi}}_{2,T}(0), \\
Q^{\boldsymbol{\pi}}_{2,T-1}(0, (1, 0)) &= 1 + V^{\boldsymbol{\pi}}_{2,T}(0), \\
Q^{\boldsymbol{\pi}}_{2,T-1}(0, (1, 1)) &= 3/4 + 1/4V^{\boldsymbol{\pi}}_{2,T}(1) + 3/4V^{\boldsymbol{\pi}}_{2,T}(0).
\end{aligned}$$

By plugging $V^{\boldsymbol{\pi}}_{i,T}(s) = 0$, we get that the policy $\boldsymbol{\pi}^*_{T-1}(0) = (1, 0)$ is the equilibrium policy, since:

$$\begin{aligned}
&Q^{\boldsymbol{\pi}}_{1,T-1}(0, (0, 0)) = -1,\ Q^{\boldsymbol{\pi}}_{2,T-1}(0, (0, 0)) = 1, \\
&Q^{\boldsymbol{\pi}}_{1,T-1}(0, (0, 1)) = -1,\ Q^{\boldsymbol{\pi}}_{2,T-1}(0, (0, 1)) = -1, \\
&Q^{\boldsymbol{\pi}}_{1,T-1}(0, (1, 0)) = 1,\ Q^{\boldsymbol{\pi}}_{2,T-1}(0, (1, 0)) = 1, \\
&Q^{\boldsymbol{\pi}}_{1,T-1}(0, (1, 1)) = -5,\ Q^{\boldsymbol{\pi}}_{2,T-1}(0, (1, 1)) = 3/4.
\end{aligned}$$

For $s = 1$, however, we have that $\boldsymbol{\pi}^*_{T-1}(1) = (0, 0)$, since:

$$\begin{aligned}
&Q^{\boldsymbol{\pi}}_{1,T-1}(1, (0, 0)) = 2,\ Q^{\boldsymbol{\pi}}_{2,T-1}(1, (0, 0)) = 1, \\
&Q^{\boldsymbol{\pi}}_{1,T-1}(1, (0, 1)) = 2,\ Q^{\boldsymbol{\pi}}_{2,T-1}(1, (0, 1)) = 0, \\
&Q^{\boldsymbol{\pi}}_{1,T-1}(1, (1, 0)) = 2,\ Q^{\boldsymbol{\pi}}_{2,T-1}(1, (1, 0)) = -1, \\
&Q^{\boldsymbol{\pi}}_{1,T-1}(1, (1, 1)) = -5,\ Q^{\boldsymbol{\pi}}_{2,T-1}(1, (1, 1)) = 3/4.
\end{aligned}$$

Therefore the V-values at time index $T - 1$ is equal to:

$$\begin{aligned}
V^{\boldsymbol{\pi}^*}_{1,T-1}(0) = 1, V^{\boldsymbol{\pi}^*}_{2,T-1}(0) = 1, \\
V^{\boldsymbol{\pi}^*}_{1,T-1}(1) = 2, V^{\boldsymbol{\pi}^*}_{2,T-1}(1) = 1.
\end{aligned}$$

We consider the test-time camouflage attack by permute $s = 0$ and $s = 1$. Therefore, assuming the agents have learned the optimal policy at time index $T - 1$, by camouflage state s, their “optimal” V-values is changed to

$$\begin{aligned}
V^{\boldsymbol{\pi}^*}_{1,T-1}(0) = -1, V^{\boldsymbol{\pi}^*}_{2,T-1}(0) = 1, \\
V^{\boldsymbol{\pi}^*}_{1,T-1}(1) = 2, V^{\boldsymbol{\pi}^*}_{2,T-1}(1) = -1,
\end{aligned}$$

which are less than the original V-values without the camouflage attack from the perspective of global benefits. At $s = 0$, after camouflage attack, the sum of V-values of agents at $T-1$ is reduced from 2 to 0 , while at $s = 1$, the sum of V-values is reduced from 3 to 1.

As for the state perception attacks, attackers can attack each agent's perception of s individually. There are totally four possible delusional joint states $s_d \in \{(0,0),(0,1),(1,0),(1,1)\}$. When $s = 0$ and $s = 1$, the attacked V-values are displayed in the following table:

V-values ↓ s_d →	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$V^{\boldsymbol{\pi}^*}_{i,T-1}(0)$	$(1,1)$	$(1,1)$	$(-1,1)$	$(-1,1)$
$V^{\boldsymbol{\pi}^*}_{i,T-1}(1)$	$(2,-1)$	$(2,-1)$	$(2,1)$	$(2,1)$

From the table we can see that the best state perception attack has the same effect as the camouflage attack in this example, where the minimal attacked global expected reward is 0 for $s = 0$ and 1 for $s = 1$.

References

[1] Shalev-Shwartz, Shaked Shammah, and Amnon Shuashua. Safe, multi-agent, reinforcement learning for autonomous driving. In *arXiv preprint arXiv:1610.03295*, 2016.

[2] Xiao-Yang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Dan Wang. Finrl: A deep reinforcement learning library for automated stock trading in quantitative finance. In *Deep RL Workshop*. NeurIPS, 2020.

[3] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, and Jiliang Tang. Deep reinforcement learning for page-wise recommendations. In *The 12th ACM Conference on Recommender Systems*, pages 95–103. ACM, 2018.

[4] Nicholas Mastronarde and Mihaela van der Schaar. Fast reinforcement learning for energy-efficient wireless communication. *IEEE Transactions on Signal Processing*, 59(12):6262–6266, 2011.

[5] Yunlong Song, Mats Steinweg, Elia Kaufmann, and Davide Scaramuzza. Autonomous drone racing with deep reinforcement learning. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, 2021.

[6] Guanlin Liu and Lifeng Lai. Provably efficient black-box action poisoning attacks against reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 12400–12410, 2021.

[7] Yanchao Sun, Ruijie Zheng, Yongyuan Liang, and Furong Huang. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep rl. In *International Conference on Learning Representations*, 2022.

[8] YenChen Lin, Zhangwei Hong, Yuanhong Liao, Mengli Shih, Mingyu liu, and Min sun. Tactics of adversarial attack on deep reinforcement learning agents. In *The 26th International Joint Conference on Artificial Intelligence*, page 3756–3762, 2017.

[9] Amin Rakhsha, Xuezhou Zhang, Xiaojin Zhu, and Adish Singla. Reward poisoning in reinforcement learning: attacks against unknown learners in unknown environments. In *arXiv preprint arXiv:2102.08492*, 2021.

[10] Yanchao Sun, Da Huo, and Furong Huang. Vulnerability-aware poisoning mechanism for online rl with unknown dynamics. In *International Conference on Learning Representations*, 2021.

[11] Kiarash Banihashem, Adish Singla, and Goran Radanovic. Defense against reward poisoning attacks in reinforcement learning. In *arXiv preprint arXiv:2102.05776*, 2021.

[12] Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Robust policy gradient against strong data corruption. In *International Conference on Machine Learning,*, pages 12391–12401, 2021.

[13] Yifang Chen, Simon S. Du, and Kevin Jamieson. Improved corruption robust algorithms for episodic reinforcement learning. In *International Conference on Machine Learning*, pages 1561–1570, 2021.

[14] Thodoris Lykouris, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245, 2021.

[15] Zun Yan, Peng Cheng, Zhuo Chen, Yonghui Li, and Branka Vucetic. Gaussian process reinforcement learning for fast opportunistic spectrum access. *IEEE Transactions on Signal Processing*, 68:2613–2628, 2020.

[16] Weitong Zhai, Xiangrong Wang, Xianbin Cao, Maria S Greco, and Fulvio Gini. Reinforcement learning based dual-functional massive mimo systems for multi-target detection and communications. *IEEE Transactions on Signal Processing*, 71:741–755, 2023.

[17] Spilios Evmorfos, Konstantinos I. Diamantaras, and Athina P. Petropulu. Reinforcement learning for motion policies in mobile relaying networks. *IEEE Transactions on Signal Processing*, 70:850–861, 2022.

[18] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *arXiv preprint arXiv:2002.04017*, 2020.

[19] Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *arXiv preprint arXiv:2010.01604*, 2020.

[20] Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *arXiv preprint arXiv:2404.18909*, 2024.

[21] Guanlin Liu and Lifeng Lai. Efficient adversarial attacks on online multi-agent reinforcement learning. In *arXiv preprint arXiv:2307.07670*, 2023.

[22] Yunhan Huang and Quanyan Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *International Conference on Decision and Game Theory for Security*, pages 217–237. Springer, 2019.

[23] Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. Adaptive reward-poisoning attacks against reinforcement learning. In *International Conference on Machine Learning*, volume 119, pages 11225–11234, 2020.

[24] Vahid Behzadan and Arslan Munir. Vunerability of deep reinforcement learning to policy induction attacks. In *International Conference on Machine Learning and Data Mining in Pattern Recogonition*, pages 262–275. Springer, 2017.

[25] Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. Policy poisoning in batch reinforcement learning and control. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[26] Michael Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference*, 1994.

[27] Hang Xu, Xinghua Qu, and Zinovi Rabinovich. Policy resilience to environment poisoning attacks on reinforcement learning. In *arXiv preprint arXiv:2304.12151*, 2023.

[28] Yunhang Huang and Quanyan Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *International Conference on Decision and Game Theory for Security*, 2019.

[29] Huan Zhang, Hongge Chen, Duane S. Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. *ArXiv*, abs/2101.08452, 2021.

[30] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations, 2021.

[31] Ziqing Lu, Guanlin Liu, Lifeng Lai, and Weiyu Xu. Optimal cost constrained adversarial attacks for multiple agent systems. In *arXiv preprint arXiv:2311.00859*, 2023.

[32] Jing Yu, Clement Gehring, Florian Schäfer, and Animashree Anandkumar. Robust reinforcement learning: A constrained game-theoretic approach. In *Proceedings of Machine Learning Research*, 2021.

[33] Sayak Mukherjee and Veronica Adetola. A secure learning control strategy via dynamic camouflaging for unknown dynamical systems under attacks. In *2021 Control Technology and Applications (CCTA),*, 2021.

[34] Wei Zhang, Qikai Zhou, Ruizhi Li, and Fu Niu. Research on camouflaged human target detection based on deep learning. In *Computational Intelligence and Neuroscience*, volume 2022, page 7703444, 2022.

[35] Xiaoqiang Tang and Bingzhe He. Malicious code dynamic traffic camouflage detection based on deep reinforcement learning in power system. In *2021 International Conference on New Energy and Power Engineering (ICNEPE 2021)*, 2021.

[36] L. Ziqing, Guanlin Liu, Lifeng Lai, and X. Weiyu. Camouflage adversarial attacks on multiple agent systems. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 7–12, 2024.