Chen, Xu; Di, Xuan; Li, Zechu

Article

Social learning for sequential driving dilemmas

Games

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Chen, Xu; Di, Xuan; Li, Zechu (2023) : Social learning for sequential driving dilemmas, Games, ISSN 2073-4336, MDPI, Basel, Vol. 14, Iss. 3, pp. 1-12, https://doi.org/10.3390/g14030041

This Version is available at:
https://hdl.handle.net/10419/330034

games

MDPI

Article

Social Learning for Sequential Driving Dilemmas

Xu Chen [1], Xuan Di [1,2,*] and Zechu Li [3]

[1] Department of Civil Engineering and Engineering Mechanics, Columbia University, New York, NY 10027, USA
[2] Data Science Institute, Columbia University, New York, NY 10027, USA
[3] Department of Computer Science, Columbia University, New York, NY 10027, USA
* Correspondence: xd2187@columbia.edu or sharon.di@columbia.edu; Tel.: +1-212-853-0435

Abstract: Autonomous driving (AV) technology has elicited discussion on social dilemmas where trade-offs between individual preferences, social norms, and collective interests may impact road safety and efficiency. In this study, we aim to identify whether social dilemmas exist in AVs' sequential decision making, which we call "sequential driving dilemmas" (SDDs). Identifying SDDs in traffic scenarios can help policymakers and AV manufacturers better understand under what circumstances SDDs arise and how to design rewards that incentivize AVs to avoid SDDs, ultimately benefiting society as a whole. To achieve this, we leverage a social learning framework, where AVs learn through interactions with random opponents, to analyze their policy learning when facing SDDs. We conduct numerical experiments on two fundamental traffic scenarios: an unsignalized intersection and a highway. We find that SDDs exist for AVs at intersections, but not on highways.

Keywords: social learning; sequential driving dilemma (SDD); autonomous vehicles (AV)

Citation: Chen, X.; Di, X; Li, Z. Social Learning for Sequential Driving Dilemmas. *Games* **2023**, *14*, 41. https://doi.org/10.3390/g14030041

Academic Editors: Heinrich H. Nax and Ulrich Berger

Received: 28 March 2023
Revised: 5 May 2023
Accepted: 10 May 2023
Published: 11 May 2023

1. Introduction

Despite significant efforts in artificial intelligence (AI) being focused on computer vision, the intelligence of autonomous vehicles (AVs) lies in their optimal decision-making abilities in the motion-planning stage while driving alongside human drivers [1–3]. The question of how individual AVs should develop the ability to survive in a complex traffic environment is a prerequisite to realizing the anticipated societal benefits of AVs [4,5], such as improved traffic safety [6–9], efficiency [10–13], and the promotion of sustainable human–machine ecologies. One major advantage that AVs have over human drivers is their ability to promptly assess situations with a greater amount of information, allowing them to react in an optimal way.

However, AVs still lack the ability to deal with sequential decision-making when facing complex driving scenarios, which may result in underperformance compared to humans. To elaborate, the self-driving technology remains unclear about how to handle the conflicts between individual self-interest and the collective interest of a group, the so-called "social dilemma". Social dilemmas can arise when individual AVs prioritize their own safety and efficiency, potentially compromising the safety and efficiency of other AVs or human-driven vehicles on the road. For instance, when changing lanes on highways, AVs must consider the speed and trajectory of surrounding vehicles while prioritizing their own utility. If each AV prioritizes its self-interest, it may lead to car accidents and decreased the overall efficiency on the road. Understanding social dilemmas involving AVs is critical at the early stage of AV adoption.

The aim of this paper is to study social dilemmas that arise when AVs make sequential decisions in a dynamic driving environment (as shown in Figure 1). To achieve this, we propose a social learning framework in which AVs learn cooperation and defection strategies in a Markov game through interactions with random opponents selected from an agent pool. By obtaining the empirical payoff matrix of the Markov game using these

strategies, we can identify the specific dilemma the game belongs to. These dilemmas, which we refer to as "sequential driving dilemmas" (SDDs), can aid policymakers in creating regulations that promote safe and efficient driving behavior, as well as designing incentive mechanisms that encourage cooperative driving strategies that benefit the group as a whole.

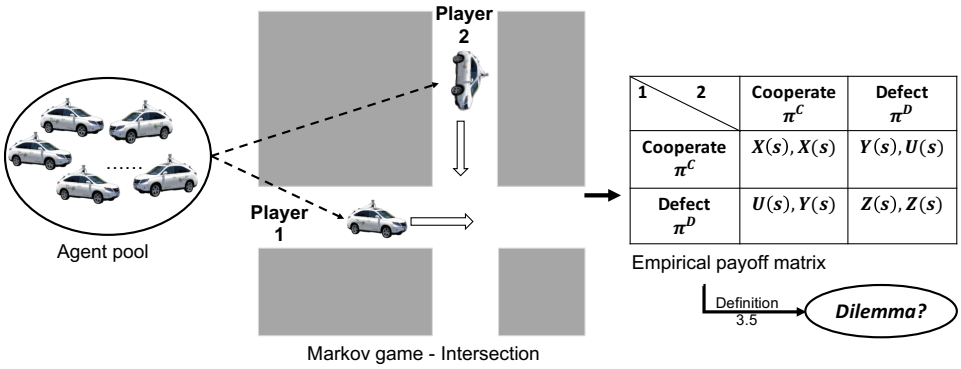

Figure 1. SDD framework (adapted from [14]).

2. Related Work

Social dilemmas have been widely studied to understand cooperative and defective behaviors among players, as seen in the works of Allen et al. [15], Alex et al. [16], Qi et al. [17], and Hilbe et al. [18]. While recent studies have investigated social dilemmas in the context of autonomous vehicles (AVs), such as the Trolley problem, which aims to investigate how AVs can make moral decisions [19,20], these studies are limited to matrix games. Matrix games are stateless and cannot capture the sequential decision-making process in dynamic traffic environments. Moreover, they do not account for other road users that may affect the state of the traffic environment. To overcome these limitations, this paper focuses on Markov games, which allow for sequential decision-making and cumulative rewards over time, making them suitable for modeling AVs' behavior in traffic environments.

To identify dilemmas in Markov games, sequential social dilemmas have been proposed to capture cooperative and defective behaviors in a dynamic environment, building on the concept of social dilemmas in two-player matrix games [14,21–23]. Prior research assumes that agents play with a fixed opponent in Markov games, which does not work for real-world scenarios in which AVs may randomly interact with each other. To tackle this challenge, this paper leverages a social learning framework where a population of agents [24] learn policies through repeated interactions with randomly selected opponents in multi-agent systems. In this study, we aim to identify sequential social dilemmas for AVs in traffic scenarios, utilizing a social learning scheme for Markov games.

Many other studies focus on enhancing coordination and cooperation among players at game equilibrium. One approach to guiding players' cooperative behavior in a social group is through social norms or conventions, which are shared standards of acceptable behavior by groups [25–28] and can lead to desired social outcomes [24]. It is important to note that social dilemmas depict a game where there is a conflict between collective and individual interests, while social norms denote a game equilibrium that can lead to desired social outcomes. We provide a summary of related work on social dilemmas and social norms in Table 1.

Table 1. Literature on social dilemmas and social norms.

	Matrix Game -Fixed Player	**Matrix Game -Social Learning**	**Markov Game -Fixed Player**	**Markov Game -Social Learning**
Social dilemma	[15–18]	—	[14,21–23,29]	This work
Social norm	[24,30–33]	[24]	—	[34]

Contributions of This Paper

The contributions of this paper include:

1. We propose a social learning framework to investigate sequential driving dilemmas (SDD) in autonomous driving systems.
2. We develop a reinforcement learning algorithm for AVs' policy learning to estimate SDDs in traffic scenarios.
3. We apply the proposed algorithm to two traffic scenarios: an unsignalized intersection and a highway, in order to identify SDDs.

The remainder of this paper is organized as follows. In Section 3, we first present preliminaries regarding social dilemmas in matrix games. In Section 4, we introduce a social learning framework to model interactions among random AV players in dynamic driving environments. In Section 5, we conduct several numerical experiments to identify SDDs in traffic scenarios. Section 6 concludes and discusses the future work.

3. Preliminaries

3.1. Social Dilemma in a Matrix Game

In this subsection, we briefly introduce social dilemmas in a traditional matrix game. Table 2 demonstrates the payoff matrix of a game between two players. Both players can choose cooperative and defective policies, which are defined as follows.

Definition 1. *Define $\pi^C \in \Pi^C$ and $\pi^D \in \Pi^D$ as cooperative and defective policies, respectively. Players choose to cooperate with opponents when conducting π^C and defect with opponents when conducting π^D. Π^C and Π^D denote policy sets.*

Table 2. Matrix game.

	π^C	π^D
π^C	X, X	Y, U
π^D	U, Y	Z, Z

Definition 2. *The game is a social dilemma if the payoff satisfies the following conditions [35]:*

1. *$X > Z$: Agents prefer mutual cooperation to mutual defection;*
2. *$X > Y$: Agents prefer mutual cooperation to unilateral cooperation;*
3. *$2X > Y + U$: Agents prefer mutual cooperation over an equal probability of unilateral cooperation and defection;*
4. *$U > X$: Agents prefer unilateral defection to mutual cooperation,*
   $Z > Y$: Agents prefer mutual defection to unilateral cooperation.

Definition 3. *According to Definition 2, we can identify three types of social dilemmas:*

1. *$U > X,\ Z > Y$: Prisoner's dilemma;*
2. *$U > X,\ Z < Y$: Chicken game;*
3. *$U < X,\ Z > Y$: Stag hunt game.*

Note that social dilemma in Definition 2 is a one-shot matrix game. The cooperative and defective policies π^C, π^D are static strategies, which are not applicable to real-world scenarios. We will extend it to sequential driving dilemmas with dynamic settings in the following subsection.

3.2. Sequential Driving Dilemma in a Markov Game

We first introduce a Markov game to model the sequential decision making of agents in a dynamic driving environment. We assume the Markov game is a non-cooperative game in which each agent aims to maximize their own cumulative payoff. The Markov game is denoted by $\mathcal{M}$. We specify each component of the Markov game as follows:

- m. There are m adaptive AVs in the agent pool, denoted by $\{1, 2, \cdots, m\}$.
- $s \in \mathcal{S}$. The environment state s in the driving environment, denoted by $s \in \mathcal{S}$, refers to global information such as the spatial distribution of all road users and road conditions. It is important to note that there may also be other road users, such as background vehicles, who are non-strategic players in the Markov game. The environment state space is denoted by $\mathcal{S}$. However, it should be noted that the environment state s is not fully observable to agents, making the Markov game a partially observable Markov decision process (POMDP).
- $o \in O$. Each agent draws a private observation from their neighborhood environment, which is a subset of the global environment state s. Specifically, agent $i \in 1, 2, \cdots, m$ draws a private observation denoted by $o_i \in O_i$, where O_i is the observation space of agent i. The joint observation space for all agents is denoted by $O = O_1 \times O_2, \times \cdots \times O_m$, which captures the overall observation of the driving environment by all agents. It is important to note that each agent is limited to only observing their surroundings and not the entire environment state.
- $a \in \mathcal{A}$. For simplicity, we adopt discrete action space. Joint action is $a = (a_1, a_2, \cdots, a_m)$, where $a_i \in A_i$, $i = 1, 2, ..., m$ and A_i is the action set. Actions in different traffic scenarios will be detailed in Section 5.
- $p \in \mathcal{P}$. After taking action a in state s, an agent arrives at a new state s' with transition probability $p(s'|s, a)$. The agent interacts with the environment to gain state transition experiences, i.e., (s, a, s').
- $r \in \mathcal{R}$. Agent $i \in \{1, 2, \cdots, m\}$ receives a reward $r_i(s, a, s')$ at each time step, which can be the travel cost in the driving environment.
- γ. The discount factor γ is used to discount the future reward. In this study, we take $\gamma = 1$, because drivers usually complete trips in a finite horizon and they value future and immediate rewards equally.

Agent $i \in \{1, 2, \cdots, m\}$ uses a policy $\pi_i : O_i \times \mathcal{A}_i \rightarrow [0, 1]$ to choose actions after drawing observation o_i. The policy is designed to maximize the agent's expected cumulative reward by selecting an optimal action. This process of observing the environment, selecting an action, and receiving a reward repeats until the agents reach their own terminal state, which occurs when either a crash happens or the agents complete their trips. In a multi-agent system, the optimal policy for agent i is denoted as π_i^*, and is defined as the agent's best response to the policies of other agents when those policies are held constant. For all agents $i \in 1, \cdots, m$, the value achieved by agent i from any state s is maximized given the other agents' policies.

$$V_i(s) = max_{a_i} \mathbb{E}_{a^{-i} \sim A} \mathbb{E}_{s' \sim P(\cdot|s,a)} [r_i(s, a, s') + \gamma V_i(s')], \tag{1}$$

where, a^{-i} is the joint action of agent i's opponents. We denote the value function of agent i starting from state s as $V_i^{\pi_i, \pi_{-i}}(s)$ given a policy π_i of agent i and policies π_{-i} of her random opponents in the environment. This study employs independent reinforcement learning to facilitate individual agents' optimal policy learning. Each agent is trained with its own policy networks.

We now present how agents learn the cooperative policy π^C and defective policy π^D in the Markov game, as well as the method for calculating the empirical payoff matrix of the agents with respect to these policies. To obtain the cooperative policy π^C and defective policy π^D within the social learning framework, we introduce a weight parameter w to represent the desired level of social outcomes [29] in an agent's reward. Specifically, we define the weight associated with desired social behaviors for the cooperative policy π^C as w^C and that for the defective policy π^D as w^D. The reward r_i obtained by agent i at each time step in a dynamic environment is then calculated as follows:

$$r_i^C = (1 - w^C) \cdot r_i^a + w^C \cdot r_i^s \tag{2}$$

$$r_i^D = (1 - w^D) \cdot r_i^a + w^D \cdot r_i^s \tag{3}$$

where, r_i^a is the reward associated with the selected action and r_i^s is the reward associated with some desired social outcome (e.g., road safety). Note that the weight parameter w is used to balance the importance of these two types of rewards. We have $w^C > w^D$, which means agents prefer collective interest to self-interest when adopting π^C and vice versa. Cooperative π^C and defective π^D policies are trained based on w^C and w^D, respectively (See the learning algorithm in Section 4.2). There are many other ways to obtain π^C and π^D. For instance, a cooperative agent is defined as a driver who always yields at the intersection, and a defective agent is a driver who crosses the intersection aggressively without taking into account other road users [36]. The level of aggressiveness is utilized for each agent via social behavior metrics [14] to determine π^C and π^D.

Definition 4. *Define an empirical payoff matrix for a pair of agents i and j who play a Markov game starting from state $s \in \mathcal{S}$ in Table 3:*

Table 3. Empirical payoff matrix in a Markov game.

	π_j^C	π_j^D
π_i^C	$X(s), X(s)$	$Y(s), U(s)$
π_i^D	$U(s), Y(s)$	$Z(s), Z(s)$

Each cell in the matrix denotes the payoff regarding π^C and π^D. Mathematically,

$$\begin{aligned} X(s) &= V_i^{\pi_i^C, \pi_j^C}(s) = V_j^{\pi_i^C, \pi_j^C}(s), \\ Z(s) &= V_i^{\pi_i^D, \pi_j^D}(s) = V_j^{\pi_i^D, \pi_j^D}(s), \\ Y(s) &= V_i^{\pi_i^C, \pi_j^D}(s) = V_j^{\pi_i^D, \pi_j^C}(s), \\ U(s) &= V_i^{\pi_i^D, \pi_j^C}(s) = V_j^{\pi_i^C, \pi_j^D}(s), \end{aligned} \tag{4}$$

where, $V_i^{\pi_i, \pi_j}(s) = \mathbb{E}_{a_t \sim (\pi_i, \pi_j)} \mathbb{E}_{s_{t+1} \sim P(\cdot|s_t, a_t)} [\sum_{t=0}^{T} r_{t+1} | s_0 = s]$. $V_i^{\pi_i, \pi_j}(s)$ is the payoff of agent i when agents i and j adopt π_i and π_j, respectively.

Definition 5. *A Markov game is a sequential driving dilemma (SDD) when there exist states $s \in \mathcal{S}$ for a pair of agents whose empirical payoff matrix satisfies the following social dilemma conditions (Definition 2):*

1. *$X(s) > Z(s)$: Agents prefer mutual cooperation to mutual defection when starting from state $s \in \mathcal{S}$;*
2. *$X(s) > Y(s)$: Agents prefer mutual cooperation to unilateral cooperation when starting from state $s \in \mathcal{S}$;*
3. *$2X(s) > Y(s) + U(s)$: When starting from state $s \in \mathcal{S}$, agents prefer mutual cooperation over an equal probability of unilateral cooperation and defection;*

4. $U(s) > X(s)$: *When starting from state* $s \in \mathcal{S}$*, agents prefer unilateral defection to mutual cooperation,*
   $Z(s) > Y(s)$: *Agents prefer mutual defection to unilateral cooperation.*

Note that the empirical payoff matrix in a Markov game is utilized to identify the existence of dilemmas. We will investigate game equilibrium in sequential driving dilemmas in the future.

4. Social Learning Framework

In this section, we introduce a social learning framework that utilizes a Markov game to model interactions among autonomous vehicles (AVs) that are randomly selected from an agent pool.

4.1. Social Learning Scheme

In order to understand road users' behaviors in the traffic environment, each AV must learn through interactions with its opponents. In traditional learning schemes, an agent repeatedly interacts with a fixed opponent until a stable equilibrium is reached. However, in the context of navigating the traffic environment, AVs encounter different opponents dynamically instead of a fixed one.

In a social learning scheme, agents learn through interactions with opponents randomly selected from an agent pool. As a result, each agent plays games repeatedly with random opponents and develops its own policy based on personal experience. This approach allows AVs to learn from a diverse set of opponents and adapt to changing circumstances, which is critical for ensuring safe and efficient driving behavior in real-world scenarios. In contrast to the fixed agents in traditional learning schemes, the agents in a social learning scheme are randomly selected from a pool in each episode. These agents then play a Markov game and update their policies based on the outcomes.

4.2. Social Learning Algorithm

In this section, we introduce a multi-agent reinforcement learning algorithm based on Deep Q-network (DQN) [37] to learn the cooperative policy π^C and defective policy π^D for AVs in the social learning scheme. The proposed algorithm is summarized in Algorithm 1.

We first initialize deep Q-networks for each agent. Q-networks for cooperative and defective policies are denoted as $Q^C(o, a; \theta^C)$, $Q^D(o, a; \theta^D)$ parameterized by θ^C and θ^D, respectively. Their target networks are $\tilde{Q}^C(o, a; \tilde{\theta}^C)$ and $\tilde{Q}^D(o, a; \tilde{\theta}^D)$. Hyperparameters, including exploration rate ϵ, learning rate η, update period τ and update parameter δ, are predetermined (See Table 4). For each run in one episode, two agents are randomly selected and removed till the agent pool is empty. When an agent is selected to participate in a Markov game, she randomly chooses to either cooperate or defect with her opponent. If agent i chooses cooperation, her cooperative Q-network $Q_i^C(o, a; \theta^C)$ and target cooperative Q-network $\tilde{Q}_i^C(o, a; \tilde{\theta}^C)$ will be updated with the weight parameter w^C of rewards in the Markov game. Similarly, if agent i chooses defection, her defective Q-network $Q_i^D(o, a; \theta^D)$ and target defective Q-network $\tilde{Q}_i^D(o, a; \tilde{\theta}^D)$ will be updated with the weight parameter w^D of rewards in the Markov game. From an initial environmental state s_0, agent i draws private observation o_i and taking action a_i according to the widely used ϵ-greedy method (i.e., agent i chooses action randomly with probability ϵ and greedily from optimal policy π_i^* with probability $1 - \epsilon$), until reaching some terminal state. The joint action $\mathbf{a}$ is executed in the environment, which results in a state transition $s \to s'$. After the state transition, each agent i receives a new private observation o_i' and a corresponding reward r_i according to Equation (3). This process is repeated until a terminal state is reached, which occurs when a crash happens or agents complete their trips. The experience tuple (o_i, a_i, r_i, o_i') of agent i is stored in a replay buffer, which is used to update the target network every τ time steps.

Algorithm 1 DQN-SDD

Initialize two networks for cooperative and defective policies π^C and π^D for each agent in the agent pool: $Q^C(o,a;\theta^C)$, $Q^D(o,a;\theta^D)$ and their target networks $\tilde{Q}^C(o,a;\tilde{\theta}^C)$, $\tilde{Q}^D(o,a;\tilde{\theta}^D)$.
Input: exploration parameter ϵ, learning rates η, target network update period τ and update parameter δ.
for $episode \leftarrow 1$ to T **do**
    while the agent pool is not empty **do**
        Randomly select two agents from the pool to play a Markov game.
        Each agent can choose to cooperate or defect with her opponent.
        while s is not terminal **do**
            For each agent, select action using ϵ-greedy policy;
            Update state transition $s \rightarrow s'$;
            Store (o_i, a_i, r_i, o_i') for each agent.
        end while
        Decay ϵ.
        if $episode > \tau$ **then**
            Update target networks: $\tilde{\theta}_i \leftarrow (1-\delta)\tilde{\theta}_i + \delta\theta_i$.
        end if
    end while
    Update Q-function with the selected learning rate η.
    Refill the pool with agents who have updated policies.
end for

After the training process, the trained Q-networks are utilized to calculate the empirical payoff matrix. First, each pair of agents from the agent pool is randomly selected. Then, the pair executes four Markov games, corresponding to the combinations of cooperative and defective policies, namely (π_C, π_C), (π_C, π_D), (π_D, π_C), and (π_D, π_D), according to their own Q-networks $Q^C(o,a;\theta^C)$ and $Q^D(o,a;\theta^D)$. The empirical payoff for agent i is calculated as cumulative rewards received over a fixed number of iterations. Each cell in the resulting payoff matrix is calculated based on Equation (4), which represents the average reward obtained by agent i when the pair of agents takes actions **a**, given the current state s and policies π^C and π^D.

Table 4. Hyperparameters.

Hyperparameter	Value
Optimizer	Adam
Number of hidden layers	3
Dropout rate	0.2
Activation function	ReLU
Learning rate η	0.0001
Initial epsilon ϵ	1.0
Epsilon decay	0.01
Final epsilon	0.01
Replay buffer size	200
Update period τ	50
Update parameter δ	0.2
Number of training episodes T	20,000

5. Numerical Experiments

In this section, we investigate two traffic scenarios (Figure 2): unsignalized intersections and highways. In each traffic scenario, we first introduce the environment set-up and then discuss SDD results.

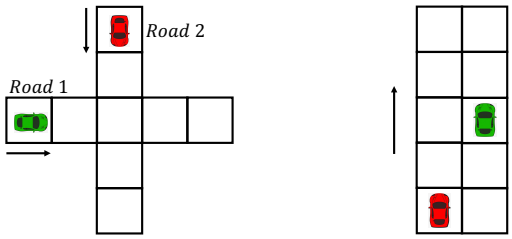

Figure 2. Traffic scenarios. (**a**) Unsignalized intersection. (**b**) Highway.

5.1. Scenario 1: Unsignalized Intersection

5.1.1. Environment Set-Up

The traffic environment is an unsignalized intersection (Figure 2a) comprising two intersecting roads, Road 1 and Road 2. Vehicles represented by green and red colors navigate Road 1 and Road 2, respectively. The intersection is discretized into uniform cells, and the state s is defined by the locations of the agents. Each agent is able to observe only the road she navigates, and the action set consists of two actions: "Go" and "Stop". The "Go" action corresponds to moving forward by one cell, while the "Stop" action represents taking no action. In each time step, agents receive a negative reward for taking either the "Go" or "Stop" action, denoting the instantaneous travel time, in order to encourage them to complete the game quickly. If a collision occurs at the intersection cell, each agent incurs a negative reward to reflect the cost of the collision. On the other hand, if no collision occurs, agents complete their trips and the game terminates. The desired social outcome is that no collision occurs, and agents successfully complete their trips.

5.1.2. SDD Results

We investigate whether SDD exists at the unsignalized intersection. In Figure 3, each point represents the empirical payoff matrix (Table 3) obtained by a pair of agents randomly selected from the pool and play the intersection game from some initial state s. For example, to estimate one cell (π_i^C, π_j^C) in the payoff matrix, we simulate the game between these two agents driven by policies π_i^C and π_j^C for 1000 times and compute the average payoff. The x-axis represents $Z(s) - Y(s)$ and the y-axis represents $U(s) - X(s)$ in the empirical payoff matrix. The squares and circles represent cases satisfying and violating SDD conditions, respectively. According to Definition 4, the intersection game is an SDD. The cooperation and defection performance in two scenarios are summarized in Table 5.

Table 5. Cooperation and defection.

	Unsignalized Intersection	**Highway**
Cooperation	one vehicle yields to another	vehicles merge and form a car platoon
Defection	vehicles do not yield to each other	vehicles stay in their own lanes

We look into points satisfying SDD conditions (orange squares). Note that these points are in the first quadrant where $U(s) > X(s)$ and $Z(s) > Y(s)$. According to Definition 4, $U(s) > X(s)$ means agents prefer unilateral defection to mutual cooperation. It implies that when AVs encounter those who choose to yield at the intersection, they prefer crossing the intersection aggressively over yielding to others. $Z(s) > Y(s)$ means agents prefer mutual defection to unilateral cooperation. It implies that when AVs encounter aggressive agents, they prefer crossing the intersection aggressively over yielding to others. According to Definition 3, the intersection game is a prisoner's dilemma.

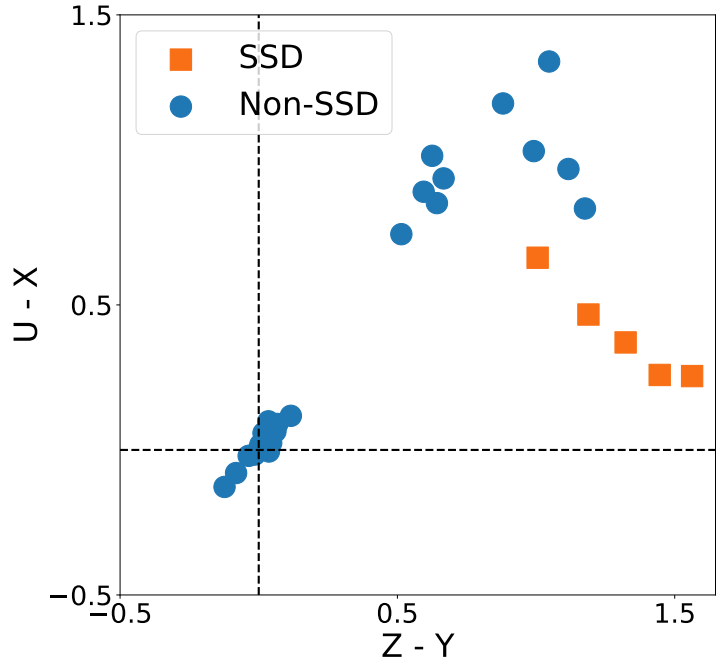

Figure 3. SDD—unsignalized intersection.

5.2. Scenario 2: Highway

5.2.1. Environment Set-Up

The traffic environment is a highway (Figure 2b) with two lanes, going from south to north. Vehicles in green and red colors represent AV players in a Markov game. The environment is discretized into uniform cells, and the state space $\mathcal{S}$ is defined by the locations of the agents. The initial locations of the agents on two lanes are randomized. Agents are capable of observing both lanes. The action set comprises three actions, namely, "Go", "Stop", and "Lane-change". The "Go" action is associated with moving forward by one cell, whereas the "Stop" action is associated with taking no action. The "Lane-change" action allows switching to the other lane. At each time step, agents receive a negative reward that denotes their travel time for taking any of the three actions. Agents receive a positive reward for formulating a car platoon, which refers to a group of vehicles that travel in a coordinated manner, with one leading vehicle and the others following behind [38].

5.2.2. SDD Results

Figure 4 displays the empirical payoff matrices obtained by simulating the platoon game. The points in the plot are located in the fourth quadrant where $U(s) < X(s)$ and $Z(s) < Y(s)$, indicating that the platoon game does not satisfy the SDD condition. Specifically, $U(s) < X(s)$ implies that agents prefer mutual cooperation over unilateral defection, indicating that AVs are more likely to form a platoon when interacting with cooperative agents. On the other hand, $Z(s) < Y(s)$ implies that agents prefer unilateral cooperation over mutual defection, suggesting that AVs are more likely to change lanes and form a platoon with defective agents.

Summarizing the two games, we find that the intersection game is an SDD but the platoon game is not. The intuitive explanation is: In the intersection game, the collective interest for agents is to avoid crashes at the intersection and individual interest is to minimize travel time. If agents want to avoid crashes, they need to stop and wait till one of them crosses the intersection, which increases travel time. It is shown that there exists a conflict between the collective and individual interests. In the presence of defective agents, the intersection game becomes a Prisoner's dilemma where agents may defect to gain individual advantages at the cost of collective interest.

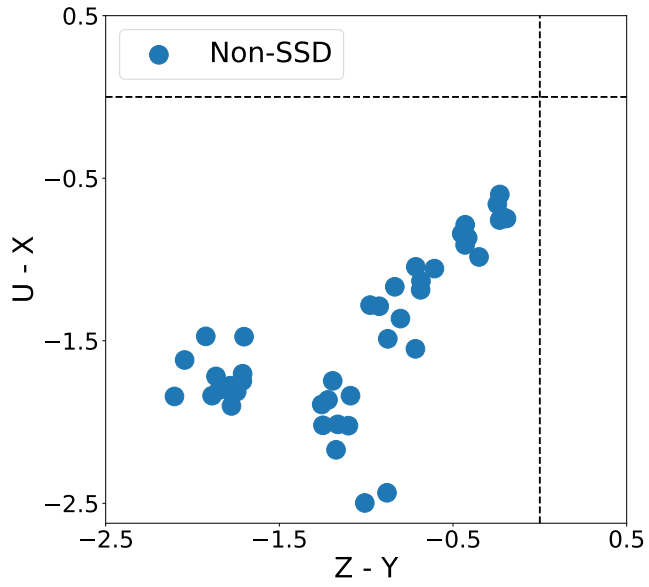

Figure 4. SDD—highway.

6. Conclusions

In this study, we employ a social learning scheme for Markov games to identify SDDs in AVs' policy learning. We propose a learning algorithm to train cooperative and defective policies and evaluate SDDs in traffic scenarios. We investigate the existence of SDDs in intersection and platoon games. The overall findings include: (1) The intersection game is an SDD, while the platoon game on highways is not. (2) The Markov game at an unsignalized intersection resembles a Prisoner's dilemma, where agents may defect to gain individual advantages, but at the cost of the collective interest.

We briefly discuss the limitations of this work: (1) AVs' decision making in driving environments are simplified as discretized action sets. There are many continuous decision variables for AV players, including velocity, brake rate and headway. (2) The number of agents who randomly encounter in Markov games is limited. The number of agents and their interactions in real-world traffic scenarios can be larger and more complex.

This work can be extended in the following ways: (1) We will leverage multi-agent reinforcement learning (MARL) and identify SDDS in more complex real-world scenarios with many road users (e.g., pedestrians). (2) Addressing SDDs for AVs will require policy-makers and road planners to design policies and incentives that encourage AVs to make decisions that benefit the public good. This may involve the creation of regulations that enforce safe and efficient behavior, in line with established traffic rules and social norms. Incentive mechanisms such as gifting in a multi-agent system can be explored to study how to enhance AVs' cooperative behaviors.

Author Contributions: Conceptualization, X.C. and X.D.; methodology, X.C.; validation, X.C. and Z.L.; writing—original draft preparation, X.C.; writing—review and editing, X.D.; visualization, X.C. and Z.L.; supervision, X.D.; project administration, X.D.; funding acquisition, X.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the National Science Foundation CAREER under award number CMMI-1943998.

Data Availability Statement: Not applicable.

Conflicts of Interest: We confirm that neither the manuscript nor any parts of its content are currently under consideration or published in another journal.

References

1. Sadigh, D.; Sastry, S.; Seshia, S.A.; Dragan, A.D. Planning for Autonomous Cars that Leverage Effects on Human Actions. In Proceedings of the Robotics: Science and Systems, Ann Arbor, MI, USA, 12–16 July 2016.
2. Fisac, J.F.; Bronstein, E.; Stefansson, E.; Sadigh, D.; Sastry, S.S.; Dragan, A.D. Hierarchical Game-Theoretic Planning for Autonomous Vehicles. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada 20–24 May 2019; pp. 9590–9596. [CrossRef]
3. Di, X.; Shi, R. A survey on autonomous vehicle control in the era of mixed-autonomy: From physics-based to AI-guided driving policy learning. *Transp. Res. Part Emerg. Technol.* **2021**, *125*, 103008. [CrossRef]
4. Huang, K.; Chen, X.; Di, X.; Du, Q. Dynamic driving and routing games for autonomous vehicles on networks: A mean field game approach. *Transp. Res. Part Emerg. Technol.* **2021**, *128*, 103189. [CrossRef]
5. Shou, Z.; Chen, X.; Fu, Y.; Di, X. Multi-agent reinforcement learning for Markov routing games: A new modeling paradigm for dynamic traffic assignment. *Transp. Res. Part Emerg. Technol.* **2022**, *137*, 103560. [CrossRef]
6. Pedersen, P.A. *A Game Theoretical Approach to Road Safety*; Technical Report, Department of Economics Discussion Paper; University of Kent: Canterbury, UK, 2001.
7. Pedersen, P.A. Moral hazard in traffic games. *J. Transp. Econ. Policy (JTEP)* **2003**, *37*, 47–68.
8. Chatterjee, I.; Davis, G. Evolutionary game theoretic approach to rear-end events on congested freeway. *Transp. Res. Rec. J. Transp. Res. Board* **2013**, *2386*, 121–127. [CrossRef]
9. Chatterjee, I. Understanding Driver Contributions to Rear-End Crashes on Congested Freeways and Their Implications for Future Safety Measures. PhD Thesis, University of Minnesota, Minneapolis, MN, USA, 2016.
10. Yoo, J.H.; Langari, R. Stackelberg game based model of highway driving. In Proceedings of the ASME 2012 5th Annual Dynamic Systems and Control Conference joint with the JSME 2012 11th Motion and Vibration Conference, Fort Lauderdale, FL, USA, 17–19 October 2012; American Society of Mechanical Engineers: New York, NY, USA, 2012; pp. 499–508.
11. Yoo, J.H. A Game Theory Based Model of Human Driving with Application to Autonomous and Mixed Driving. Doctoral Dissertation, Texas A & M University, College Station, TX, USA, 2014.
12. Talebpour, A.; Mahmassani, H.; Hamdar, S. Modeling Lane-Changing Behavior in a Connected Environment: A Game Theory Approach. *Transp. Res. Part Emerg. Technol.* **2015**, *59*, 216–232. [CrossRef]
13. Yu, H.; Tseng, H.E.; Langari, R. A human-like game theory-based controller for automatic lane changing. *Transp. Res. Part Emerg. Technol.* **2018**, *88*, 140–158. [CrossRef]
14. Leibo, J.Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; Graepel, T. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In Proceedings of the AAMAS '17, 16th International Conference on Autonomous Agents and MultiAgent Systems, Sao Paulo, Brazil, 8–12 May 2017.
15. Allen, B.; Lippner, G.; Chen, Y.T.; Fotouhi, B.; Momeni, N.; Yau, S.T.; Nowak, M.A. Evolutionary dynamics on any population structure. *Nature* **2017**, *544*, 227–230. [CrossRef]
16. McAvoy, A.; Allen, B.; Nowak, M.A. Social goods dilemmas in heterogeneous societies. *Nat. Hum. Behav.* **2020**, *4*, 819–831. [CrossRef]
17. Su, Q.; McAvoy, A.; Wang, L.; Nowak, M.A. Evolutionary dynamics with game transitions. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 25398–25404. [CrossRef]
18. Hilbe, C.; Štěpán Š.; Chatterjee, K.; Nowak, M.A. Evolution of cooperation in stochastic games. *Nature* **2018**, *559*, 246–249. [CrossRef] [PubMed]
19. Bonnefon, J.F.; Shariff, A.; Rahwan, I. The social dilemma of autonomous vehicles. *Science* **2016**, *352*, 1573–1576. [CrossRef] [PubMed]
20. Schwarting, W.; Pierson, A.; Alonso-Mora, J.; Karaman, S.; Rus, D. Social behavior for autonomous vehicles. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 24972–24978. [CrossRef] [PubMed]
21. Eccles, T.; Hughes, E.; Kramár, J.; Wheelwright, S.; Leibo, J.Z. Learning Reciprocity in Complex Sequential Social Dilemmas. *arXiv* **2019**, arXiv:1903.08082.
22. Badjatiya, P.; Sarkar, M.; Sinha, A.; Singh, S.; Puri, N.; Subramanian, J.; Krishnamurthy, B. Inducing Cooperative behaviour in Sequential-Social dilemmas through Multi-Agent Reinforcement Learning using Status-Quo Loss. *arXiv* **2020**, arXiv:2001.05458.
23. Gupta, G. Obedience-Based Multi-Agent Cooperation for Sequential Social Dilemmas. Master Thesis, University of Waterloo, Waterloo, ON, Canada, 2020.
24. Sen, S.; Airiau, S. Emergence of Norms through Social Learning. In Proceedings of the IJCAI'07, 20th International Joint Conference on Artifical Intelligence, Hyderabad, India, 6–12 January 2007; Morgan Kaufmann Publishers Inc: San Francisco, CA, USA, 2007; pp. 1507–1512.
25. Lewis, D. *Convention: A Philosophical Study*; Wiley: Hoboken, NJ, USA, 1970.
26. Boella, G.; Lesmo, L. *A Game Theoretic Approach to Norms and Agents*; Universita di Torino: Torino, Italy, 2001.
27. Boella, G.; van der Torre, L. Norm governed multiagent systems: The delegation of control to autonomous agents. In Proceedings of the IAT 2003, IEEE/WIC International Conference on Intelligent Agent Technology, Halifax, NSA, Canada, 13–17 October 2003; pp. 329–335. [CrossRef]
28. Epstein, J. Learning to Be Thoughtless: Social Norms and Individual Computation. *Comput. Econ.* **2001**, *18*, 9–24. [CrossRef]

29. O'Callaghan, D.; Mannion, P. *Tunable Behaviours in Sequential Social Dilemmas Using Multi-Objective Reinforcement Learning*; International Foundation for Autonomous Agents and Multiagent Systems: Richland, SC, USA, 2021; pp. 1610–1612.
30. Delgado, J. Emergence of social conventions in complex networks. *Artif. Intell.* **2002**, *141*, 171–185. [CrossRef]
31. Villatoro, D.; Sabater-Mir, J.; Sen, S. Social Instruments for Robust Convention Emergence. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011; pp. 420–425. [CrossRef]
32. Yu, C.; Zhang, M.; Ren, F.; Luo, X. Emergence of Social Norms through Collective Learning in Networked Agent Societies. In Proceedings of the AAMAS '13, 2013 International Conference on Autonomous Agents and Multi-Agent Systems, St. Paul, MN, USA, 6–10 May 2013; International Foundation for Autonomous Agents and Multiagent Systems: Richland, SC, USA, 2013; pp. 475–482.
33. Franks, H.; Griffiths, N.; Jhumka, A. Manipulating convention emergence using influencer agents. *Auton. Agents-Multi-Agent Syst.* **2013**, *26*, 315–353. [CrossRef]
34. Chen, X.; Li, Z.; Di, X. Social Learning In Markov Games: Empowering Autonomous Driving. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 4–9 June 2022; pp. 478–483. [CrossRef]
35. Macy, D.; Flache, A. Learning Dynamics in Social Dilemmas. *Proc. Natl. Acad. Sci. USA* **2002**, *99* (Suppl. 3), 7229–7236. [CrossRef]
36. Bouderba, S.; Moussa, N. Evolutionary dilemma game for conflict resolution at unsignalized traffic intersection. *Int. J. Mod. Phys.* **2019**, *30*, 189. [CrossRef]
37. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.A.; Fidjeland, A.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
38. Chen, X.; Di, X. Legal Framework for Rear-End Crashes in Mixed-Traffic Platooning: A Matrix Game Approach. *Future Transp.* **2023**, *3*, 417–428. [CrossRef]