Emergent Prosociality in Multi-Agent Games Through Gifting

Woodrow Z. Wang[1*], **Mark Beliaev**[2*], **Erdem Bıyık**[1*], **Daniel A. Lazar**[2], **Ramtin Pedarsani**[2] and **Dorsa Sadigh**[1]

[1]Stanford University, [2]University of California, Santa Barbara

{wwang153, ebiyik, dorsa}@stanford.edu, {mbeliaev, dlazar, ramtin}@ucsb.edu

Abstract

Coordination is often critical to forming prosocial behaviors – behaviors that increase the overall sum of rewards received by all agents in a multi-agent game. However, state of the art reinforcement learning algorithms often suffer from converging to socially less desirable equilibria when multiple equilibria exist. Previous works address this challenge with explicit reward shaping, which requires the strong assumption that agents can be forced to be prosocial. We propose using a less restrictive peer-rewarding mechanism, *gifting*, that guides the agents toward more socially desirable equilibria while allowing agents to remain selfish and decentralized. Gifting allows each agent to give some of their reward to other agents. We employ a theoretical framework that captures the benefit of gifting in converging to the prosocial equilibrium by characterizing the equilibria's basins of attraction in a dynamical system. With gifting, we demonstrate increased convergence of high risk, general-sum coordination games to the prosocial equilibrium both via numerical analysis and experiments.

1 Introduction

Reinforcement learning (RL) has shown great success in training agents to solve many human-relevant tasks [Mnih *et al.*, 2013; Sutton *et al.*, 1999]. In addition, there has been increased interest in leveraging RL techniques in decentralized multi-agent problems, motivated by outstanding performance in two-player zero-sum games such as AlphaZero [Silver *et al.*, 2017]. However, simply applying multi-agent RL algorithms to train self-interested agents in a decentralized fashion does not always perform well. Specifically, win-win strategies – strategies that are beneficial for all agents – are often challenging to achieve in more advanced settings, such as general-sum games where win-win outcomes are only possible through coordination [Matignon *et al.*, 2012].

Coordination is often coupled with risk. In the real world, there are many applications where there is a safe action that leads to guaranteed but lower rewards, and a risky action that leads to higher rewards only if agents cooperate, such as alleviating traffic congestion [Bıyık *et al.*, 2018; Lazar *et al.*, 2019], sustainably sharing limited resources [Hughes *et al.*, 2018], and altruistic human-robot interaction [Shirado and Christakis, 2017]. In a game-theoretic framework, the class of Stag Hunt games is a well-known instance of the trade-off between social cooperation and safety. In Stag Hunt, two players must independently decide between foraging mushrooms and hunting a stag. If both players choose to hunt the stag, they succeed and are rewarded greatly, while if only one of them goes hunting, they return empty-handed and injured. On the other hand, foraging mushrooms guarantees a meal for the night, although not as satisfying (as shown in Fig. 1). When confronted with this coordination problem, state of the art RL algorithms and humans alike often choose the safer, socially less desirable option of foraging, instead of the riskier, prosocial option of hunting [Van Huyck *et al.*, 1990; Peysakhovich and Lerer, 2018]. This is because uncertainty in the behavior of other players leads one to favor safer alternatives. Ideally, one would prefer reaching the most socially desirable equilibrium – the *prosocial equilibrium*[1] – to allow all agents to maximize their rewards.

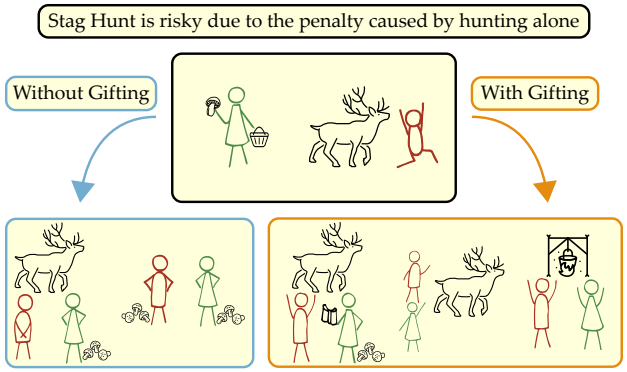

Figure 1: Progression of agents in a Stag Hunt game with and without gifting. Agents choose either to hunt or forage. Hunting provides more food, but requires coordination: an agent is severely penalized for hunting alone. Foraging guarantees a small amount of food. Without gifting, agents often learn to forage. With gifting, they gift each other early in training, which mitigates risk, and they learn to hunt together in a prosocial manner.

*First three authors have contributed equally and listed randomly.

[1]When multiple equilibria are equally prosocial, we refer to reaching any of them as reaching the prosocial equilibrium.

Previous attempts to address the problem of reaching the prosocial equilibrium focus on using explicit reward shaping to force agents to be prosocial, such as by making agents care about the rewards of their partners [Peysakhovich and Lerer, 2018]. This requires the strong assumption that a central agent, e.g., a supervisor, can coerce the agents to be altruistic and to care about maximizing the total social utility. We are interested in a less restrictive setting where we do not assume access to a centralized supervisor, and the agents retain their self-interest and only care about maximizing their own received reward – while the objective is still to increase the probability of reaching the prosocial equilibrium.

Our key insight is that gifting, a peer-rewarding mechanism, can be used during training as a decentralized teaching mechanism, mitigating risk and encouraging prosocial behavior without explicit reward shaping. Gifting was first introduced by Lupu and Precup [2020] and is an instance of a larger class of algorithms that extend an agent's action space, providing them with a way to give part of their reward to other agents through their actions. In contrast to centralized reward shaping, which requires an external actor to force agents to be prosocial apriori, gifting leaves it up to the agents themselves to use the new actions, enabling self-interested agents to decide *when* and *how* to use the gifting actions.

One key advantage of gifting is that it can be used at training time as a behavior shaping mechanism by allowing agents to take risk-mitigating actions, i.e., agent 1 can gift agent 2 in order to decrease agent 2's risk and incentivize agent 2 to take a riskier action. We prove zero-sum gifting does not introduce new pure-strategy Nash equilibria in one-shot normal-form games. On the other hand, gifting can introduce new, complex learned behaviors in a repeated normal-form game: new equilibria may be introduced where one agent's policy is contingent on receiving a gift. In this paper, we only demonstrate preliminary results in repeated normal-form games, and focus our analysis on one-shot normal-form games to carefully examine the effects of gifting as a transient risk-mitigating action used only at train time.

Our main contributions in this paper are as follows:

- We propose using a zero-sum gifting mechanism to encourage prosocial behavior in coordination games while allowing agents to remain decentralized and self-interested.
- We provide insights on the effects of zero-sum gifting for N-player one-shot normal-form games by formally showing it does not introduce new equilibria and characterizing conditions under which gifting is beneficial.
- We experimentally show that zero-sum gifting increases the probability of convergence of selfish agents to the prosocial equilibrium in general-sum coordination games when the agents are trained with Deep Q-Networks [Mnih *et al.*, 2013].

2 Related Work

Game Theory: There has been significant recent work attempting to reach the prosocial equilibrium in coordination games. Several works use tools from both multi-agent RL and game theory to investigate multi-agent learning in cooperative games. Balcan *et al.* [2015] study learning cooperative games from a set of coalition samples. More related to our work, Panait *et al.* [2008] investigate the effect of leniency, an agent's tolerance of other agents' non-cooperative actions, from an evolutionary game theory perspective. In our work, we investigate non-cooperative games in the scope of gifting. Gifting allows agents to take a new action that may lower the risk their opponent experiences, whereas leniency allows agents to ignore rewards they received in the past.

Coordinated Exploration and Centralized Training: In multi-agent games, researchers have tried coordinating the exploration of agents to find the most prosocial Nash equilibrium [Iqbal and Sha, 2019] or other equilibrium concepts [Beliaev *et al.*, 2020]. While coordinated exploration improves exploration efficiency, it requires communication among agents, which is not available in decentralized settings. Centralized training methods with decentralized control have also been proposed as a way to learn multi-agent policies [Lowe *et al.*, 2017; Foerster *et al.*, 2018b]. However, similar to coordinated exploration, these approaches require communication among agents during training, which is often not applicable in practice.

Opponent Modeling: Reasoning about opponent behavior can lead to more complex interactions. With opponent modeling, agents can estimate their opponents' policies, parameters, or updates in order to inform their own learning [Foerster *et al.*, 2018a; Sadigh *et al.*, 2018; Sadigh *et al.*, 2016; Shih *et al.*, 2021; Xie *et al.*, 2020; Zhu *et al.*, 2020]. While opponent modeling has shown promising results, it often provides approximation solutions that can be suboptimal. Letcher *et al.* [2019] have shown that many opponent modeling methods might prevent convergence to equilibria.

Explicit Reward Shaping: To encourage coordination, researchers have explicitly shaped the reward of agents, such as by encoding inequity aversion [Hughes *et al.*, 2018]. Peysakhovich and Lerer [2018] define each agent's reward function to be the sum of all agents' rewards in the environment. Although successful, these approaches require the strong assumption that an agent's reward function can be externally modified and that the agent can be forced to be prosocial and care about maximizing the total utility of all agents.

Gifting: Gifting is a recently proposed method that extends the action space of learning agents to allow rewards to be transferred among agents [Lupu and Precup, 2020]. It simply extends each agent's action space with gifting actions, but does not require that the agents use the new gifting actions in any particular way. In our work, we leverage the idea of gifting for improving coordination in general-sum games and examine the effects of the added gifting actions both analytically and experimentally.

3 Problem Definition

We are interested in developing and analyzing algorithms that encourage agents to exhibit prosocial behavior in multi-agent environments with multiple equilibria. We formalize this problem for general-sum one-shot coordination games.

General-sum one-shot coordination games are a class of

games with multiple pure-strategy Nash equilibria. *Pure-strategy Nash equilibria* (PNE) are game-theoretic solution concepts, in which each agent has no incentive to unilaterally deviate from a deterministic strategy given the strategy of the other agents. When applying RL techniques to these games, multi-agent systems reach one of the PNE if the agents converge to deterministic policies [Harsanyi and Selten, 2001], although not always the best PNE for all agents. Ideally, they would converge to the payoff-dominant PNE, in which at least one player receives a strictly higher payoff and no player would receive a higher payoff in another equilibrium [Harsanyi and Selten, 2001]. If such an equilibrium exists, then it is *prosocial* because the sum of rewards for all agents is larger than that of any other equilibrium.

In practice, we are interested in reaching the prosocial equilibrium; however, this is not trivial in settings such as coordination games, where some of the equilibria are risk-dominant, i.e., they have lower but more guaranteed payoffs even if the other agents do not coordinate. These equilibria have larger basins of attraction, so uncertainty in other players' behaviors would lead one to choose the risk-dominant strategy [Harsanyi and Selten, 2001]. In settings where both payoff-dominant and risk-dominant equilibria exist, it is difficult to reach the prosocial equilibrium, as agents must be willing to take risks and cooperate with each other. Stag Hunt games are an excellent example of such a setting. Many recent multi-agent RL works have studied Stag Hunt games in depth [Peysakhovich and Lerer, 2018; Nica *et al.*, 2017; Leibo *et al.*, 2017], as well as works that attempt to build AI systems coordinating with humans [Shum *et al.*, 2019].

The payoff matrix for a general two-action, two-player game is shown in Table 1.

	Action 1	Action 2
Action 1	a, A	b, B
Action 2	c, C	d, D

Table 1: Payoff matrix of a general two-player game

In a coordination game, multiple PNE exist, and they occur when players *coordinate* by choosing the same action. This restricts the PNE to lie on the main diagonal of the payoff matrix. Formally, in coordination games, we have $a > c, A > B, d > b, D > C$ [Harsanyi and Selten, 2001]. These inequalities place the PNE on the main diagonal, satisfying the condition that agents must coordinate on the same action in order to reach a PNE. Furthermore, (Action 1, Action 1) is the payoff-dominant equilibrium if $a \geq d, A \geq D$, and at least one of the inequalities is strict. The specific values of these payoffs will determine what sub-class of coordination games is being played: Pure Coordination, Bach or Stravinsky (BoS), Assurance, or Stag Hunt. We are most interested in the Stag Hunt setting because of the difficulty it presents in reaching the prosocial equilibrium. Detailed descriptions of the other sub-class games are in the Appendix.

3.1 Stag Hunt

- $a > d, A > D$
- $a = A, d = D, c = B, C = b = r$
- $a - c < d - r$

	Hunt	Forage
Hunt	$2, 2$	$r, 1$
Forage	$1, r$	$1, 1$

Stag Hunt is a two-player game with a risk-dominant equilibrium at (Forage, Forage) and a payoff-dominant equilibrium at (Hunt, Hunt). The payoff-dominant equilibrium is more prosocial and provides each agent with higher reward, but contains risk in the case where the agents do not coordinate. The risk-dominant equilibrium is safer, since the reward is less contingent on the other agent's cooperation.

A Nash equilibrium risk dominates another if it has a strictly higher Nash product [Harsanyi and Selten, 2001]. The Nash product of an equilibrium is the product of deviation losses of both players. As shown in the third condition above ($a - c < d - r$), all of the parameters influence risk when comparing Nash products of the equilibria. However, for simplicity, we characterize the risk in Stag Hunt with the parameter r, the reward for hunting alone, and we keep our analysis focused on r while holding all other values in the payoff matrix constant. In our setting of Stag Hunt, the equilibrium at (Hunt, Hunt) has a Nash product of $(2-1)^2 = 1$, while the equilibrium at (Forage, Forage) has a Nash product of $(1-r)^2$. With r strictly negative, as r decreases, the risk monotonically increases since the Nash product of (Forage, Forage) grows larger. Thus, we refer to r as the *risk-varying parameter*. In our analysis, we use three versions of Stag Hunt with $r = -2, -6, -10$, referred to as the *low*, *medium*, and *high* risk settings, respectively. If the corresponding setting is not mentioned, then we default to medium risk.

3.2 Zero-Sum Gifting

To increase the probability of reaching the prosocial equilibrium in settings with multiple equilibria, we investigate adding zero-sum gifting actions based on the work by Lupu and Precup [2020]. With zero-sum gifting, an agent may decide to give some of its reward to the others, preserving the total reward of agents. Although our main interest is with coordination games, the method can be generally applied to all normal-form games and is formalized below.

Take any normal-form game M with a finite set of N players, each with a set of actions (strategies) S_i, and payoff function $\mu_i : S_1 \times S_2 \times \ldots \times S_N \to \mathbb{R}$ where $i \in \{1, 2, \ldots, N\}$. We denote the subset $\mathbf{S_{PNE}} \subseteq S_1 \times S_2 \times \ldots \times S_N$ as the set of PNE actions in M:

$$\begin{aligned} &\boldsymbol{s} \in \mathbf{S_{PNE}} \text{ if and only if} \\ &\forall i \in \{1, 2, \ldots, N\} \text{ and } \forall s_i' \in S_i : \mu_i(\boldsymbol{s}) \geq \mu_i(s_i', \boldsymbol{s}_{-i}), \end{aligned} \tag{1}$$

where $\boldsymbol{s}_{-i}$ denotes the set of actions of all agents other than agent i. To introduce gifting, we define a new finite set of actions G_i, and function $\sigma_i : G_1 \times G_2 \times \ldots \times G_N \to \mathbb{R}$ for each player where:

$$\begin{aligned} &0 \in G_i \ , \\ &\forall g_i \in G_i : g_i \geq 0 \ , \\ &\sigma_i(\boldsymbol{g}) = -g_i + \frac{1}{N-1} \sum_{j \in -i} g_j \ . \end{aligned} \tag{2}$$

Here, σ_i formulates how the payoff of agent i changes by the gifting actions of all agents, $\boldsymbol{g}$.

We then formulate the new game $\bar{M}$ with gifting actions.

In $\bar{M}$, the set of actions for each player is $\bar{S}_i = S_i \times G_i$, and the corresponding payoffs functions are $\bar{\mu}_i : \bar{S}_1 \times \bar{S}_2 \times \ldots \times \bar{S}_N \to \mathbb{R}$ where:

$$\begin{aligned} &\forall \bar{\boldsymbol{s}} \in \bar{S}_1 \times \bar{S}_2 \times \ldots \times \bar{S}_N \text{ and } \forall i \in \{1, 2, \ldots, N\} : \\ &\bar{\mu}_i(\bar{\boldsymbol{s}}) = \mu_i(\boldsymbol{s}) + \sigma_i(\boldsymbol{g}) \text{ where } \bar{\boldsymbol{s}} = (\boldsymbol{s}, \boldsymbol{g}) \,. \end{aligned} \tag{3}$$

Since $\sum_{i=1}^{N} \sigma_i(\boldsymbol{g}) = 0$, introducing the gifting actions into the game does not change the total reward among all agents.

Having formalized zero-sum gifting for any normal-form game, we now proceed with analysis and experiments to highlight its benefits. We focus our experiments on settings where we add zero-sum gifting actions with each $G_i = \{0, \gamma\}$.

4 Analysis of Zero-Sum Gifting in One-Shot Normal-Form Games

We analyze the effect of zero-sum gifting on the equilibria of one-shot normal-form games in Section 4.1. In Section 4.2, we characterize the behavior of learning agents in Stag Hunt with gifting. Specifically, we formulate the learning process of the agents as a dynamical system and show that gifting increases the basin of attraction of the prosocial equilibrium.

4.1 Effects of Gifting on Equilibria

In this section, we state our main theoretical results. We provide the proofs for both Lemma 1 and Proposition 1 in the Appendix. We show that agents gift each other 0 reward in the PNE of $\bar{M}$. Moreover, $\bar{\mathbf{S}}_{\mathbf{PNE}}$, the set of PNE in $\bar{M}$, has a one-to-one correspondence with $\mathbf{S}_{\mathbf{PNE}}$, the PNE in the original game M. Together, these imply having gifting actions does not change the equilibrium behavior of the agents.

Lemma 1. *In any one-shot normal-form game extended with zero-sum gifting actions and for any $s_i \in S_i$, (s_i, g_i) is strictly dominated by $(s_i, 0)$ if $g_i \neq 0$, meaning $(s_i, 0)$ always leads to higher payoff for agent i than (s_i, g_i) for any action profile $\bar{\boldsymbol{s}}_{-i}$ by other agents.*

Corollary 1. *In the set of PNE of any normal-form game extended with zero-sum gifting actions, $\bar{\mathbf{S}}_{\mathbf{PNE}}$, all agents gift 0 reward.*

$$\begin{aligned} &\forall \bar{\boldsymbol{s}} \in \bar{\mathbf{S}}_{\mathbf{PNE}} \text{ and } \forall i \in \{1, 2, \ldots, N\} : \\ &g_i = 0 \text{ where } \bar{\boldsymbol{s}} = (\boldsymbol{s}, \boldsymbol{g}) \end{aligned} \tag{4}$$

Proposition 1. *For any normal-form game M extended to $\bar{M}$ with zero-sum gifting, there exists a unique one-to-one mapping between their corresponding sets of PNE strategy profiles $\mathbf{S}_{\mathbf{PNE}}$ and $\bar{\mathbf{S}}_{\mathbf{PNE}}$, such that if an action set is a PNE in M, then appending 0-gifting actions gives a PNE in $\bar{M}$:*

$$\begin{aligned} &\bigtimes_{i=1}^{N} s_i \in \mathbf{S}_{\mathbf{PNE}} \iff \bigtimes_{i=1}^{N} (s_i, 0) \in \bar{\mathbf{S}}_{\mathbf{PNE}}, \text{ and} \\ &\bigtimes_{i=1}^{N} (s_i, g_i) \in \bar{\mathbf{S}}_{\mathbf{PNE}} \implies \forall i \in \{1, 2, \ldots, N\} : g_i = 0 \,. \end{aligned} \tag{5}$$

Proposition 1 is a desirable result, because it means that introducing gifting actions to one-shot normal-form games will not change the final behavior of the learning agents in the equilibria. Thus, we can carefully investigate gifting's effect on reaching the original equilibria of the game. Moreover, we can view these extended actions as transient to the environment – the gifting actions are only seen at training time.

While not changing the final equilibrium behavior, introducing gifting actions increases the frequency of converging to a more desirable PNE in a dynamic learning environment: prosocial behavior is observed more often after agents converge to an equilibrium. Hence, it is reasonable to extend agents' action spaces with gifting actions specifically in scenarios with higher risk, as we can promote prosocial behavior without directly shaping rewards in a centralized manner.

4.2 Effects of Gifting on the Agents' Behavior

Since Stag Hunt is the most interesting and difficult game among the games we introduced in Section 3, we now analyze the behavior of learning agents in Stag Hunt with zero-sum gifting. By Proposition 1, we know they will converge to either (Hunt, Hunt) or (Forage, Forage) at PNE[2] – they will not give gifts. In this section, we analyze the deciding factors that lead agents to a specific PNE.

Our idea is to formulate the game and the learning process as a dynamical system. To do this, we first define the policies of the two agents, $\pi_{\boldsymbol{x}}$ and $\pi_{\boldsymbol{y}}$, respectively. Here, $\boldsymbol{x}$ and $\boldsymbol{y}$ parameterize the policies. While the policies may have various forms, such as neural networks, what is important is how we define the probability of each action. For that, we let $\boldsymbol{x}$ and $\boldsymbol{y}$ be the logits of a softmax policy:

$$\pi_{\boldsymbol{x}}(s^{(j)}) = \frac{\exp(x_j)}{\sum_{j'=1}^{4} \exp(x_{j'})} \,, \tag{6}$$

and similarly for $\pi_{\boldsymbol{y}}$, where $s^{(1)}$ = Hunt, $s^{(2)}$ = Forage, $s^{(3)}$ = Hunt + Gift, $s^{(4)}$ = Forage + Gift. As an example, here $\boldsymbol{x}$ and $\boldsymbol{y}$ can be the outputs of a neural network. It can be noted that only the relative differences of the parameters are important: adding a scalar to all parameters of an agent does not change the policy. The two PNEs of the game then correspond to $\forall i \in \{2, 3, 4\} : x_1 - x_i = y_1 - y_i = +\infty$ for the prosocial and $\forall i \in \{1, 3, 4\} : x_2 - x_i = y_2 - y_i = +\infty$ for the risk-dominant equilibrium.

We then write the expected reward of the agents as:

$$\mathbb{E}[\bar{\mu}_i] = \sum_{\bar{s}_1 \in \bar{S}_1} \sum_{\bar{s}_2 \in \bar{S}_2} \pi_{\boldsymbol{x}}(\bar{s}_1) \pi_{\boldsymbol{y}}(\bar{s}_2) \bar{\mu}_i(\bar{s}_1, \bar{s}_2) \,. \tag{7}$$

While RL methods employ various methods to estimate value functions or policy gradients, true gradients can be closely estimated here by collecting large batches of data at every learning iteration, as this is a one-shot game. Thus, we use the true gradients:

$$\frac{\partial \mathbb{E}[\bar{\mu}_i]}{\partial x_j} = \sum_{\bar{s}_1 \in \bar{S}_1} \sum_{\bar{s}_2 \in \bar{S}_2} \frac{\partial \pi_{\boldsymbol{x}}(\bar{s}_1)}{\partial x_j} \pi_{\boldsymbol{y}}(\bar{s}_2) \bar{\mu}_i(\bar{s}_1, \bar{s}_2) \,, \tag{8}$$

and similarly for $\frac{\partial \mathbb{E}[\bar{\mu}_i]}{\partial y_j}$. Since both agents are only self-

[2] This game also has a mixed-strategy Nash equilibrium. However, that equilibrium is unstable, so learning agents do not converge there in practice. Thus, we exclude the analysis of this equilibrium.

interested and want to maximize their reward, they will update their policies following these gradients.

This formulation leads to an 8-dimensional autonomous dynamical system (system with no input) with state $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$, and

$$\dot{\boldsymbol{z}} = f(\boldsymbol{z}) = \left[\frac{\partial \mathbb{E}[\bar{\mu}_1]}{\partial \boldsymbol{x}^\top}, \frac{\partial \mathbb{E}[\bar{\mu}_2]}{\partial \boldsymbol{y}^\top}\right]^\top \tag{9}$$

We visualize the phase portraits of this dynamical system in Appendix D.

While our formulation so far in this section is general and can be applied to any 2-player 4-action (including gifting) normal-form games, we now focus on Stag Hunt. For the remainder of this section, we take $a = A = 2$, $r = b = C = -6$, $B = c = 1$, $d = D = 1$ (the same payoff matrix as in Section 3.1 with medium risk) and gift value $\gamma = 10$. Similarly, we formulate the original game without gifting as a dynamical system.

We then want to compute the basins of attraction of the equilibria, i.e., the initial states of the system that lead to that specific equilibrium. This is possible, because we already know the stable equilibria of the dynamical systems – they are the PNEs of the corresponding normal-form games.

To this end, Fig. 2 shows the ratio of initial states that reach the prosocial equilibrium when the relative differences of gifting parameters with respect to x_2 and y_2 are taken as uniformly spaced values in $[-3, 3]$[3]. The left heatmap shows the two basins of attraction without gifting. Since the gifting parameters are irrelevant in this setting, the map is binary. The right heatmap corresponds to the extended game with zero-sum gifting actions. The blue line shows the boundary for the game without gifting for comparison. Overall, this shows gifting is indeed beneficial for getting prosocial behavior in the equilibrium.

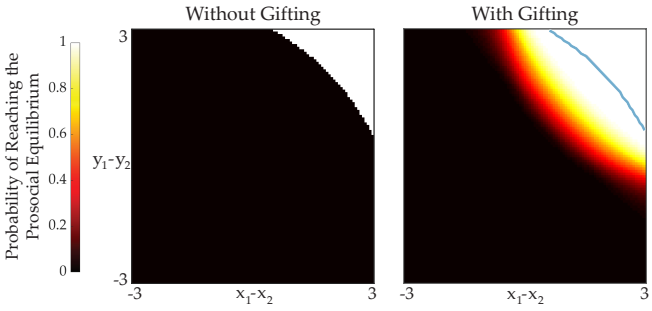

Figure 2: Heatmaps show the probability of reaching the prosocial equilibrium when the policy parameters associated with gifting ($x_3 - x_2, x_4 - x_2, y_3 - y_2, y_4 - y_2$) are taken uniformly from $[-3, 3]$. The blue curve in the right heatmap shows the boundary without gifting for comparison.

By repeating the same analysis of basin of attraction for varying $\gamma \in \{1, 2, \ldots, 20\}$ and $r \in \{-10, -6, -2\}$, we analyze how often the system converges to the prosocial equilibrium under different conditions. Fig. 3 suggests that while gifting is always helpful, its benefits become more significant when agents are allowed to gift higher amounts.

[3]As only the differences between parameters are important, we vary the gifting parameters with respect to x_2 and y_2.

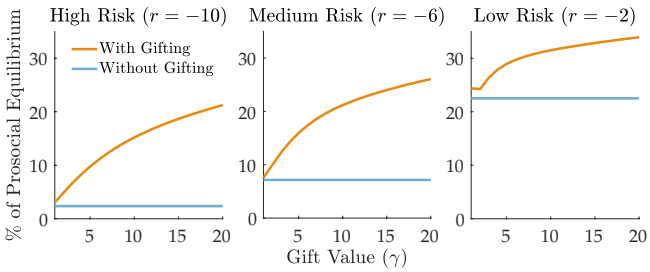

Figure 3: Frequency of reaching the prosocial equilibrium under varying risk and gift γ amounts. The relative differences of gifting parameters with respect to x_2 and y_2 are taken as uniformly spaced values in $[-3, 3]$ to compute the frequencies.

In the next section, we validate these results, which we obtained by assuming access to the true gradients, in more realistic learning settings where the payoffs are unknown to agents.

5 Experiments

We first describe the environments used in our experiments and implementation details. We then present experiment results to support our analysis.

5.1 Environments

Two-Player Environments: We test the effect of gifting in multiple popular game-theoretic coordination games: Pure Coordination, Bach or Stravinsky, Assurance, and Stag Hunt.

We perform the majority of our analysis on Stag Hunt (low, medium, and high risk), as it contains both the payoff and risk-dominated equilibria. Assurance is similar to Stag Hunt but the payoff-dominant equilibrium is no longer risk-dominated. BoS and Pure Coordination have two equally prosocial PNE, so they are included in our experiments to demonstrate our method still reaches a PNE in those settings.

N-Player Environments: We run experiments on Stag Hunt with more than two players, where the game is defined by a graph (similar to [Peysakhovich and Lerer, 2018]). Each node in the graph represents an agent and the edges define the individual Stag Hunts to be played. Thus, each agent chooses an action to play with all neighbor agents and receives the average reward of the games. We specifically examine three-player and four-player fully connected graphs in the medium risk setting: FC-3 Stag Hunt and FC-4 Stag Hunt.

Repeated Games: We further investigate the effect of zero sum gifting in repeated interactions by running experiments on a Repeated Stag Hunt environment. In Repeated Stag Hunt, agents repeatedly play the medium risk, one-shot Stag Hunt over a finite horizon $H = 10$. Each agent observes the most recent action taken by the other agent. This setting is interesting as agents have an expanded policy space: their policies are conditioned on other players' previous actions, and so new Nash equilibria may emerge with gifting.

5.2 Implementation Details

We use the payoff matrices shown in Section 3. Unless otherwise stated, we set $\gamma = 10$. For all experiments, we train a Deep Q-Network (DQN) with independent ϵ-greedy exploration for each agent. We use Adam optimizer with a learning

rate of 5×10^{-4}. The replay buffer size is 10^5. The ϵ for exploration begins at 0.3 and exponentially decays to 0.01 over 2×10^4 steps. Each target network updates every 250 episodes. For the one-shot games, all agents are given a constant observation of 0. We provide supplementary code for reproducibility of all the experiments.

Environment	*Without Gifting*	*With Gifting*
Bach or Stravinsky	**100.0%**	**100.0%**
Pure Coordination	**100.0%**	**100.0%**
Assurance	56.8%	**63.3%**
High Risk Stag Hunt	0.0%	**19.0%**
Med. Risk Stag Hunt	8.6%	**21.4%**
Low Risk Stag Hunt	**25.4%**	22.0%
FC-3 Stag Hunt	7.8%	**12.1%**
FC-4 Stag Hunt	5.1%	**7.4%**
Repeated Stag Hunt	0.0%	**19.7%**

Table 2: The percentage of **1024** runs with random initializations that reached the prosocial equilibrium with multi-agent DQN.

5.3 Results

As shown in Table 2, zero-sum gifting increases the probability of converging to the most prosocial equilibrium in a variety of coordination games. In BoS and Pure Coordination, all equilibria are equally prosocial, and we converge to one of the equilibria 100% of the time with and without gifting. In Assurance, the lack of risk makes the prosocial equilibrium a favourable outcome even without gifting, but we still see an improvement when adding gifting actions to the agents. In Stag Hunt we see that gifting has a greater benefit when risk is higher, but performance diminishes slightly in the low risk setting when gifting is introduced. This interdependence between varying risk and gifting is further explored later in this section. In the FC-3 and FC-4 Stag Hunts, we can see that gifting helps increase the probability of convergence to the prosocial equilibrium, but as the number of agents increases, it becomes more difficult to coordinate all agents and encourage risky prosocial behavior over safer actions. In Repeated Stag Hunt, gifting significantly increases the probability of convergence to the prosocial equilibrium. When compared to the results of the corresponding one-shot medium risk Stag Hunt, we can see the likelihood of agents coordinating at the prosocial equilibrium decreases both with and without gifting, implying that coordination over repeated instances of Stag Hunt is a more difficult, risky setting.

Interdependence of Risk and Gift Value: In Fig. 4, we examine the relation between the gift value and the risk value in Stag Hunt. The results show that in order for gifting to help in coordination games with higher risk, the gift value needs to increase to compensate for the added risk.

We can also see a slight negative effect gifting has in low risk settings when training with DQN agents. One explanation for this is that adding gifting actions to agents expands their action space and makes exploration more difficult, and uncertainty in the other agent's actions favors the risk-dominant equilibrium. However, under high risk settings, agents are more likely to behave prosocially with gifting, even with increased difficulty in exploration.

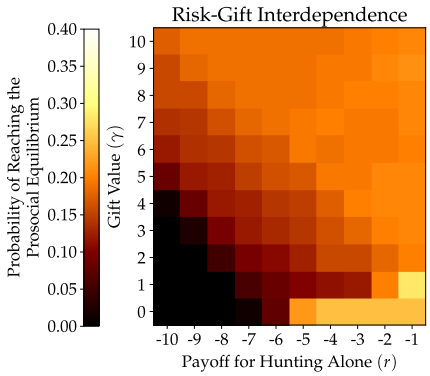

Figure 4: This figure depicts the relationship between the risk and gift value in Stag Hunt. In Stag Hunt, the payoff for hunting alone ($\mathbf{r}$) characterizes the risk. The results show that as risk increases, the gift value γ must increase proportionally in order to be risk-mitigating and improve convergence to the prosocial equilibrium.

6 Discussion

Summary: We formalize a zero-sum gifting mechanism and show that it often increases the probability of convergence to the prosocial equilibrium in coordination games. We prove that zero-sum gifting does not alter the behavior under Nash equilibria in one-shot normal-form games. With gifting, we show via numerical analysis that the prosocial equilibrium's basin of attraction grows in Stag Hunt and empirically validate these results with DQN in a broader set of environments.

Limitations: We analyze gifting as an alternative method for encouraging prosocial behavior compared to explicit reward shaping. In practice, gifting requires the ability to extend an environment's action space, so it can only be applied in settings where agents' action spaces can be modified.

Moreover, although our experimental results in Table 2 show that gifting negatively affects the low risk Stag Hunt setting when trained with DQN, the performance loss is marginal compared to the performance gain we see in higher risk settings. Nonetheless, one should be cautious when applying gifting, as the benefits are dependent on the risk in the respective environment.

Future Work: We focus the majority of our experiments on one-shot games, since we are interested in isolating the setting where no new equilibria are introduced by the gifting actions. We provide brief experiments of gifting in the repeated game setting, but further exploring the emergence of complex behaviors involving gifting in repeated interactions can help shed light on what settings gifting would be most beneficial.

Acknowledgments

We would like to thank NSF EPCN grant #1952920 and the DARPA HiCon-Learn project for their support.

References

[Balcan *et al.*, 2015] Maria-Florina Balcan, Ariel D Procaccia, and Yair Zick. Learning cooperative games. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 475–481, 2015.

[Beliaev *et al.*, 2020] Mark Beliaev, Woodrow Z. Wang, Daniel A. Lazar, Erdem Biyik, Dorsa Sadigh, and Ramtin Pedarsani. Emergent correlated equilibrium through synchronized exploration. In *RSS 2020 Workshop on Emergent Behaviors in Human-Robot Systems*, July 2020.

[Bıyık *et al.*, 2018] Erdem Bıyık, Daniel A. Lazar, Ramtin Pedarsani, and Dorsa Sadigh. Altruistic autonomy: Beating congestion on shared roads. *Algorithmic Foundations of Robotics XIII*, page 887–904, 2018.

[Foerster *et al.*, 2018a] Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130, 2018.

[Foerster *et al.*, 2018b] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Harsanyi and Selten, 2001] John Harsanyi and Reinhard Selten. *A General Theory of Equilibrium in Games*, volume 18. MIT Press, 01 2001.

[Hughes *et al.*, 2018] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*, pages 3326–3336, 2018.

[Iqbal and Sha, 2019] Shariq Iqbal and Fei Sha. Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning. *arXiv preprint arXiv:1905.12127*, 2019.

[Lazar *et al.*, 2019] Daniel A Lazar, Erdem Bıyık, Dorsa Sadigh, and Ramtin Pedarsani. Learning how to dynamically route autonomous vehicles on shared roads. *arXiv preprint arXiv:1909.03664*, 2019.

[Leibo *et al.*, 2017] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473, 2017.

[Letcher *et al.*, 2019] Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. In *International Conference on Learning Representations*, 2019.

[Lowe *et al.*, 2017] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, pages 6379–6390, 2017.

[Lupu and Precup, 2020] Andrei Lupu and Doina Precup. Gifting in multi-agent reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, 2020.

[Matignon *et al.*, 2012] Laetitia Matignon, Guillaume Laurent, and Nadine Fort-Piat. Independent reinforcement learners in cooperative markov games: A survey regarding coordination problems. *The Knowledge Engineering Review*, 27:1 – 31, 03 2012.

[Mnih *et al.*, 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In *NIPS 2013 Deep Learning Workshop*, 2013.

[Nica *et al.*, 2017] Andrei Cristian Nica, Tudor Berariu, Florin Gogianu, and Adina Magda Florea. Learning to maximize return in a stag hunt collaborative scenario through deep reinforcement learning. In *2017 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 188–195. IEEE, 2017.

[Panait *et al.*, 2008] Liviu Panait, Karl Tuyls, and Sean Luke. Theoretical advantages of lenient learners: An evolutionary game theoretic perspective. *Journal of Machine Learning Research*, 9(Mar):423–457, 2008.

[Peysakhovich and Lerer, 2018] Alexander Peysakhovich and Adam Lerer. Prosocial learning agents solve generalized stag hunts better than selfish ones. In *International Conference on Autonomous Agents and Multiagent Systems*, 2018.

[Sadigh *et al.*, 2016] Dorsa Sadigh, S. Shankar Sastry, Sanjit A. Seshia, and Anca D. Dragan. Planning for autonomous cars that leverage effects on human actions. In *Proceedings of Robotics: Science and Systems (RSS)*, June 2016.

[Sadigh *et al.*, 2018] Dorsa Sadigh, Nick Landolfi, S. Shankar Sastry, Sanjit A. Seshia, and Anca D. Dragan. Planning for cars that coordinate with people: Leveraging effects on human actions for planning and active information gathering over human internal state. *Autonomous Robots (AURO)*, 42(7):1405–1426, October 2018.

[Shih *et al.*, 2021] Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, and Dorsa Sadigh. On the critical role of conventions in adaptive human-{ai} collaboration. In *International Conference on Learning Representations*, 2021.

[Shirado and Christakis, 2017] Hirokazu Shirado and Nicholas A Christakis. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, 545(7654):370–374, 2017.

[Shum *et al.*, 2019] Michael Shum, Max Kleiman-Weiner, Michael L Littman, and Joshua B Tenenbaum. Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6163–6170, 2019.

[Silver *et al.*, 2017] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017.

[Sutton *et al.*, 1999] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

[Van Huyck *et al.*, 1990] John B Van Huyck, Raymond C Battalio, and Richard O Beil. Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure. *American Economic Review*, 80(1):234–248, March 1990.

[Xie *et al.*, 2020] Annie Xie, Dylan Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. Learning latent representations to influence multi-agent interaction. In *Proceedings of the 4th Conference on Robot Learning (CoRL)*, November 2020.

[Zhu *et al.*, 2020] Zheqing Zhu, Erdem Biyik, and Dorsa Sadigh. Multi-agent safe planning with gaussian processes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2020.

A Sub-classes of Coordination Games

We formally define each sub-class of coordination game that we provide experimental results for in Table 2. We define the conditions for each coordination game, as well as provide a concrete example in the form of a payoff matrix.

A.1 Pure Coordination

- $b = B = c = C = \zeta$
- $a = d, A = D$
- $\zeta < \min(a, A)$

	Action 1	Action 2
Action 1	$1, 1$	$0, 0$
Action 2	$0, 0$	$1, 1$

In the simplest type of coordination game, there does not exist a payoff-dominant equilibrium, as both PNE give identical payoffs and are equally prosocial. There is also no additional risk associated with choosing Action 2 as opposed to Action 1, and hence, we expect randomly initialized learning agents to converge to either equilibrium with equal probability.

A.2 Bach or Stravinsky (BoS)

- $b = B = c = C = \zeta$
- $a > d, A < D$
- $\zeta < \min(a, A)$

	Action 1	Action 2
Action 1	$2, 1$	$0, 0$
Action 2	$0, 0$	$1, 2$

In BoS, the PNE are not identical, and neither PNE is payoff-dominant. The row player prefers (Action 1, Action 1) and the column player prefers (Action 2, Action 2). When $a = D$, $A = d$, we consider either PNE to be the prosocial equilibrium, as the sum of rewards among players is identical. In this case, both equilibria have the same risk, and we expect randomly initialized learning agents to converge to either equilibrium with equal probability.

A.3 Assurance

- $b = B = c = C = \zeta$
- $a > d, A > D$
- $\zeta < \min(a, A)$

	Action 1	Action 2
Action 1	$2, 2$	$0, 0$
Action 2	$0, 0$	$1, 1$

In the game of Assurance, the payoff-dominant PNE is (Action 1, Action 1). It is also risk-dominant, i.e., even if an agent thinks their partner may not coordinate by taking the same action, there is no incentive to go for Action 2. This makes it easy for agents to reach the payoff-dominant equilibrium.

B Proof of Lemma 1

Lemma 1. *In any one-shot normal-form game extended with zero-sum gifting actions and for any $s_i \in S_i$, (s_i, g_i) is strictly dominated by $(s_i, 0)$ if $g_i \neq 0$, meaning $(s_i, 0)$ always leads to higher payoff for agent i than (s_i, g_i) for any action profile $\bar{\boldsymbol{s}}_{-i}$ by other agents.*

Proof. For any $\bar{\boldsymbol{s}}_{-i} = (\boldsymbol{s}_{-i}, \boldsymbol{g}_{-i}) \in \bar{\boldsymbol{S}}_{-i}$, the payoff for agent i under the action (s_i, g_i) is

$$\bar{\mu}_i(\bar{\boldsymbol{s}}) = \mu_i(\boldsymbol{s}) + \sigma_i(\boldsymbol{g}) = \mu_i(s) - g_i + \frac{1}{N-1} \sum_{j \in -i} g_j \tag{10}$$

If agent i had $(s_i, 0)$, its payoff would be $\mu_i(\boldsymbol{s}) + \frac{1}{N-1} \sum_{j \in -i} g_j$, which is strictly larger as $g_i > 0$, regardless of $\bar{\boldsymbol{s}}_{-i}$. Hence, (s_i, g_i) is strictly dominated by $(s_i, 0)$, and this completes the proof. □

C Proof of Proposition 1

Proposition 1. *For any normal-form game M extended to $\bar{M}$ with zero-sum gifting, there exists a unique one-to-one mapping between their corresponding sets of PNE strategy profiles $\mathbf{S_{PNE}}$ and $\bar{\mathbf{S}}_{\mathbf{PNE}}$, such that if an action set is a PNE in M, then appending 0-gifting actions gives a PNE in $\bar{M}$:*

$$\begin{aligned} &\bigtimes_{i=1}^{N} s_i \in \mathbf{S_{PNE}} \iff \bigtimes_{i=1}^{N} (s_i, 0) \in \bar{\mathbf{S}}_{\mathbf{PNE}}, \text{ and} \\ &\bigtimes_{i=1}^{N} (s_i, g_i) \in \bar{\mathbf{S}}_{\mathbf{PNE}} \implies \forall i \in \{1, 2, \ldots, N\} : g_i = 0 \, . \end{aligned} \tag{5}$$

Proof. We already know from Corollary 1 that actions involving non-zero gifting cannot exist in the PNE of $\bar{M}$, so the latter statement is true. We now prove the former statement. First, we define $\bar{\mathbf{S}}_{\mathbf{0}}$ by appending 0-gifting actions to the action sets in $\mathbf{S_{PNE}}$:

$$\bar{\mathbf{S}}_{\mathbf{0}} = \left\{ \bigtimes_{i=1}^{N} (s_i, 0) \mid \boldsymbol{s} \in \mathbf{S_{PNE}} \right\} \, . \tag{11}$$

Next, we show that $\bar{\mathbf{S}}_{\mathbf{0}} = \bar{\mathbf{S}}_{\mathbf{PNE}}$, proving the first statement.

$$\begin{aligned} &\forall \bar{\boldsymbol{s}} \in \mathbf{S_0} \text{ and } \forall i \in \{1, 2, \ldots, N\} : \\ &\bar{\mu}_i(\bar{\boldsymbol{s}}) = \mu_i(\boldsymbol{s}) + \sigma_i\left((0, 0, \ldots, 0)\right) = \mu_i(\boldsymbol{s}) + 0 \end{aligned}$$

Because $\boldsymbol{s} \in \mathbf{S_{PNE}}$, we have

$$\forall s_i' \in S_i : \mu_i(\boldsymbol{s}) \geq \mu_i(s_i', \boldsymbol{s}_{-i}) \, ,$$

implying changing s_i only does not increase the payoff for agent i. Moreover, we know from Lemma 1 that any non-zero gifting action is strictly dominated by the corresponding zero-gifting action. These two results mean changing the gifting action g_i, the original action s_i, or both cannot increase the payoff for agent i:

$$\forall \bar{s}_i' \in \bar{S}_i : \bar{\mu}(\bar{\boldsymbol{s}}) \geq \bar{\mu}(\bar{s}_i', \bar{\boldsymbol{s}}_{-i}) \, , \tag{12}$$

and therefore $\bar{\mathbf{S}}_{\mathbf{0}} = \bar{\mathbf{S}}_{\mathbf{PNE}}$. □

D Gradients of the Dynamical System

Figure 5 shows the *normalized* gradients of the system that govern the dynamics for various parameters of gifting actions (x_3, x_4, y_3 and y_4). Again, as only the differences between parameters are important, we vary the gifting parameters with respect to x_2 and y_2. Since the prosocial equilibrium is reached with $x_1 - x_2 = y_1 - y_2 = +\infty$ and the risk-dominated equilibrium with $x_1 - x_2 = y_1 - y_2 = -\infty$, these phase portraits show the two regions of states that would be updated to move towards either of the equilibria. It can be seen that higher gifting parameters enlarge the region that moves towards the prosocial equilibrium.

It should be noted that while Fig. 5 gives a picture of system dynamics, it is limited in two aspects: first, it does not provide any information about what happens when the gifting parameters are not equal to each other. Second, the gradients of individual states only give information about one-step updates learning agents would have. However, because the gifting parameters will also be learned, Fig. 5 does not show the initial states that will reach the prosocial equilibrium. Therefore, the basin of attraction analyses we made in Section 4.2 gives a more accurate picture.

E Transient Gifting Actions

As Proposition 1 shows, zero-sum gifting does not introduce any new equilibria to one-shot normal-form games. Thus, we investigate the usage of gifting actions at train time to provide insight on how gifting encourages agents to be prosocial. Fig. 6 shows that, even when agents start with frequent zero-sum gifting actions, they use gifting as transient actions during training to encourage other agents to update towards the more prosocial equilibrium. As the agents optimize their own parameters selfishly, the agents take the gifting

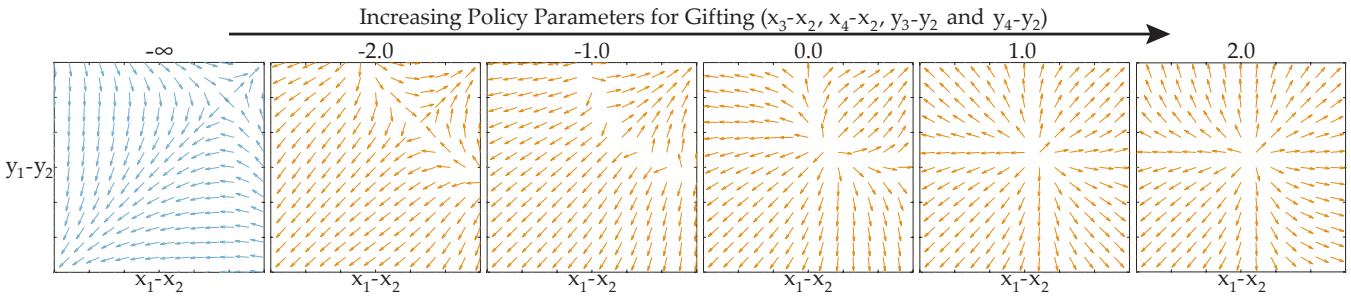

Figure 5: Phase portraits of the formulated dynamical system under various gifting parameters. Note the left-most figure shows the system of the game without gifting. Axes show $[-3, 3]$ for each plot.

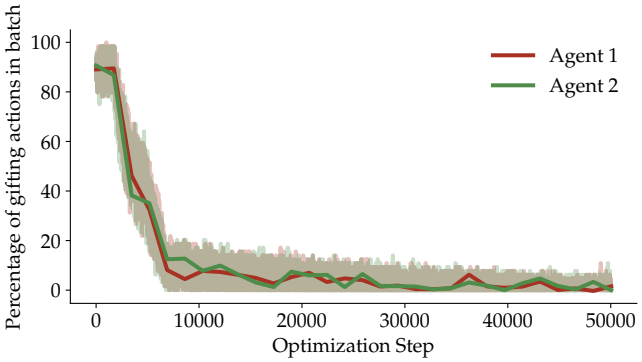

Figure 6: This plot shows the percentage of gifting actions in a batch vs. the optimization step during one training run that starts with frequent gifting actions and reaches the prosocial equilibrium in Stag Hunt. Both agents are initialized to have a higher Q-value for the gifting actions and equal Q-value for the non-gifting actions. This provides an empirical example of Lemma 1 in practice, where agents gift initially to encourage prosocial behavior, and learn not to gift in the limit, while reaching the prosocial equilibrium.

actions less frequently, and their final converged policies never include gifting actions in the case of one-shot normal-form games.

F Compute Details

The basin of attraction code ran on an Elastic Compute Cloud (EC2) instance in Amazon Web Services (AWS) with 16 vCPUs and 30 GB RAM. Each run took between 2 and 24 hours depending on how fast the agents converge to equilibria.

The DQN training code ran on a personal computer with an 8C/16T processor and 32 GB RAM. Figure 4 took 36 hours to complete. Each result in Table 2 took around 2 hours to complete.