Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning

Kevin R. McKee, Ian Gemp, Brian McWilliams,
Edgar A. Duéñez-Guzmán, Edward Hughes, & Joel Z. Leibo
DeepMind
London
{kevinrmckee,imgemp,bmcw,duenez,edwardhughes,jzl}@google.com

ABSTRACT

Recent research on reinforcement learning in pure-conflict and pure-common interest games has emphasized the importance of population heterogeneity. In contrast, studies of reinforcement learning in mixed-motive games have primarily leveraged homogeneous approaches. Given the defining characteristic of mixed-motive games—the imperfect correlation of incentives between group members—we study the effect of population heterogeneity on mixed-motive reinforcement learning. We draw on *interdependence theory* from social psychology and imbue reinforcement learning agents with Social Value Orientation (SVO), a flexible formalization of preferences over group outcome distributions. We subsequently explore the effects of diversity in SVO on populations of reinforcement learning agents in two mixed-motive Markov games. We demonstrate that heterogeneity in SVO generates meaningful and complex behavioral variation among agents similar to that suggested by interdependence theory. Empirical results in these mixed-motive dilemmas suggest agents trained in heterogeneous populations develop particularly generalized, high-performing policies relative to those trained in homogeneous populations.

1 INTRODUCTION

In multi-agent reinforcement learning, the actions of one agent can influence the experience and outcomes for other agents—that is, agents are *interdependent*. Interdependent interactions can be sorted into two categories based on the alignment of incentives for the agents involved [47]:

(1) *Pure-motive* interactions, in which the group's incentives are either entirely aligned (pure-common interest) or entirely opposed (pure-conflict),

(2) and *mixed-motive* interactions, in which the group's incentives are sometimes aligned and sometimes in conflict.[1]

Examples of the former include games such as Hanabi [6] and Go [49]. The latter category is typified by games like the Prisoner's Dilemma [46, 53] and the tragedy of the commons [20, 31]. This

---

[1]When Schelling originally introduced the pure- and mixed-motive framework, he explained, "Mixed-motive refers not, of course, to an individual's lack of clarity about his own preferences but rather to the ambivalence of his relation to the other player—the mixture of mutual dependence and conflict, of partnership and competition" [47].

categorical distinction is especially relevant for spatially and temporally extended Markov games. In these games, interdependence emerges both as a direct impact of one agent's actions on another's outcomes and as an indirect effect of each agent on the state of the substrate environment in which others co-exist.
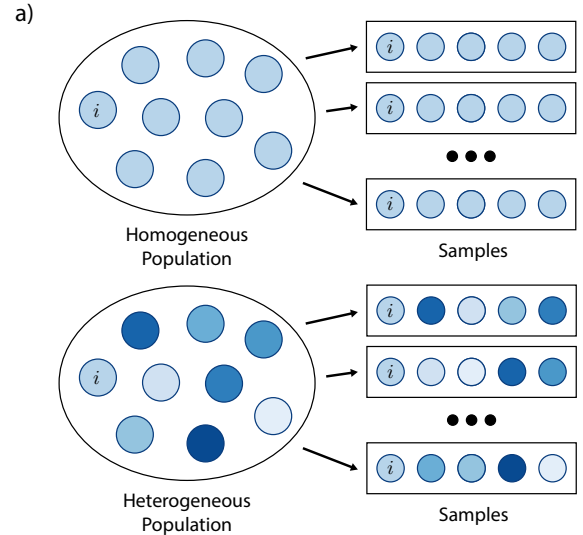
Figure 1: **Homogeneity and heterogeneity in population-based multi-agent reinforcement learning. (a) Population homogeneity and heterogeneity result in different training experiences for a given agent *i*. In the homogeneous case, agent policies are either identical or very similar (e.g., due to identical training distributions or shared motivations). In the heterogeneous setting, a given agent *i* encounters a range of group compositions over time. The variability in policies can stem from agents training under different distributions or with different motivations. (b) Representative examples of previous multi-agent reinforcement learning research. We study the mixed-motive, heterogeneous setting.**

In pure-conflict reinforcement learning, self-play solutions for Markov games [5, 22, 49] have gradually given way to population-based approaches [26, 55] (Figure 1a). A central impetus for this shift has been an interest in ensuring agent performance is robust to opponent heterogeneity (i.e., variation in the set of potential opponent policies). Similarly, recent work on pure-common interest reinforcement learning in Markov games has highlighted the importance of robustness to diverse partner policies [1, 11]. In both of these contexts, it is desirable to train agents capable of adapting and best responding to a wide range of potential policy sets.

In mixed-motive Markov games, the effects of partner heterogeneity have not received much attention. Most mixed-motive reinforcement learning research has produced policies through self-play [32, 43] or co-training policies in fixed groups [24, 31]. Such methods foster homogeneity in the set of policies each agent encounters.

We aim to introduce policy heterogeneity into mixed-motive reinforcement learning (Figure 1b). In recent years, a growing body of work has explored the effects of furnishing agents in mixed-motive games with various motivations such as inequity aversion [24, 56], social imitation [15], and social influence [28]. Thus, a natural starting point for the study of heterogeneity is to explore the effects of diversity in intrinsic motivation [51]. Here we endow our agents with Social Value Orientation (SVO), an intrinsic motivation to prefer certain group reward distributions between self and others.

Psychology and economics research has repeatedly demonstrated that human groups sustain high levels of cooperation across different games through heterogeneous distributive preferences [7, 13, 16, 45, 50]. A particularly compelling account from *interdependence*

theory holds that humans deviate from game theoretic predictions in economic games because each player acts not on the "given matrix" of a game—which reflects the extrinsic payoffs set forth by the game rules—but on an "effective matrix", which represents the set of outcomes as subjectively evaluated by that player [29]. Players receive the given matrix from their environment and subsequently apply an "outcome transformation" reflecting their individual preferences for various outcome distributions. The combination of the given matrix and the outcome transformation form the effective matrix (Figure 2). Though an individual's social preferences may decrease their given payoff in a single game, groups with diverse sets of preferences are capable of resisting suboptimal Nash equilibria.

In multi-agent reinforcement learning, reward sharing is commonly used to resolve mixed-motive dilemmas [24, 44, 52, 56]. To date, agent hyperparameters controlling reward mixture have typically been shared. This approach implies a homogeneous population of policies, echoing the representative agent assumption from economics [21, 30]. The continued reliance on shared reward mixtures is especially striking considering that the ability to capture heterogeneity is a key strength of agent-based models over other modeling approaches [19].

Homogeneous populations often fall prey to a peculiar variant of the lazy agent problem [52], wherein one or more agents begin ignoring the individual learning task at hand and instead optimize for the shared reward [24]. These shared-reward agents shoulder the burden of prosocial work, in a manner invariant to radical shifts in group composition across episodes. This "specialization" represents a failure of training to generate generalized policies.

To investigate the effects of heterogeneity in mixed-motive reinforcement learning, we introduce a novel, generalized mechanism for reward sharing. We derive this reward-sharing mechanism, Social Value Orientation (SVO), from studies of human cooperative behavior in social psychology. We show that across several games, heterogeneous distributions of these social preferences within groups generate more generalized individual policies than do homogeneous distributions. We subsequently explore how heterogeneity in SVO sustains positive group outcomes. In doing so, we demonstrate that this formalization of social preferences leads agents to discover specific prosocial behaviors relevant to each environment.
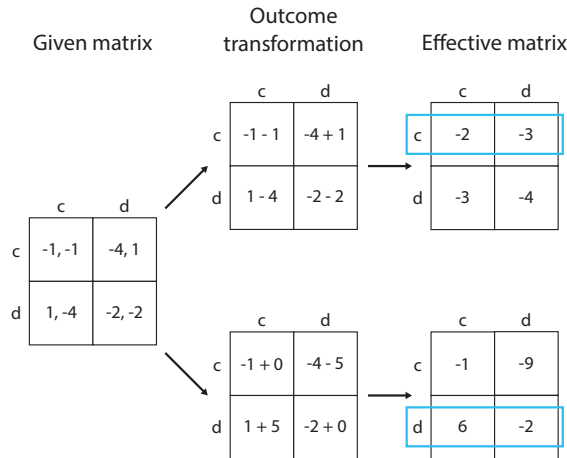

Figure 2: Interdependence theory. The top pathway depicts a transformation process for a row player who has altruistic preferences. In this case, the outcome transformation directly transfers the column player's payoff into the effective matrix. The bottom pathway shows a transformation process for a row player with competitive preferences, who finds it rewarding to maximize the distance between their payoff and the column player's payoff. These two outcome transformations suggest different dominant strategies (highlighted in blue).

2 AGENTS

2.1 Multi-agent reinforcement learning and Markov games

In this work, we consider n-player partially observable Markov games. A partially observable Markov game $\mathcal{M}$ is defined on a finite set of states $\mathcal{S}$. The game is endowed with an observation function $O : \mathcal{S} \times \{1, \ldots, n\} \rightarrow \mathbb{R}^d$; a set of available actions for each player, $\mathcal{A}_1, \ldots, \mathcal{A}_n$; and a stochastic transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \rightarrow \Delta(\mathcal{S})$, which maps from the joint actions taken by the n players to the set of discrete probability distributions over states. From each state, players take a joint action $\vec{a} = (a_1, \ldots, a_n) \in \mathcal{A}_1, \ldots, \mathcal{A}_n$.

Each agent i independently experiences the environment and learns a behavior policy $\pi(a_i|o_i)$ based on its own observation $o_i = O(s, i)$ and (scalar) extrinsic reward $r_i(s, \vec{a})$. Each agent learns

a policy which maximizes a long term γ-discounted payoff defined as:

$$V_{\vec{\pi}_i}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t U_i(s_t, \vec{o}_t, \vec{a}_t)|\vec{a}_t \sim \vec{\pi}_t, s_{t+1} \sim \mathcal{T}(s_t, \vec{a}_t)\right] \quad (1)$$

where $U_i(s_t, \vec{o}_t, \vec{a}_t)$ is a utility function and $\vec{o} = (o_1, \ldots, o_n)$ for simplicity. In standard reinforcement learning, the utility function maps directly to the extrinsic reward provided by the environment.

2.2 Social Value Orientation

Here we introduce Social Value Orientation (SVO), an intrinsic motivation to prefer certain group reward distributions between self and others.

We introduce the concept of a *reward angle* as a scalar representation of the observed distribution of reward between player i and all other players in the group. The size of the angle formed by these two scalars represents the relative distribution of reward between self and others (Figure 3). Given a group of size n, its corresponding reward vector is $\vec{R} = (r_1, \ldots, r_n)$. The reward angle for player i is:

$$\theta(\vec{R}) = \text{atan}\left(\frac{\bar{r}_{-i}}{r_i}\right) \quad (2)$$

where $\bar{r}_{-i}$ is a statistic summarizing the rewards of all other group members. We choose the arithmetic mean $\bar{r}_{-i} = \frac{1}{n-1}\sum_{j \neq i} r_j$ [38]. Note that reward angles are invariant to the scalar magnitude of $\vec{R}$.

The Social Value Orientation, θ^{SVO}, for player i is player i's target distribution of reward among group members. We use the difference between the observed reward angle and the targeted SVO to calculate an intrinsic reward. Combining this intrinsic reward with the extrinsic reward signal from the environment, we can define the following utility function U_i to be maximized in Eq. (1):

$$U_i(s, o_i, a_i) = r_i - w \cdot |\theta^{\text{SVO}} - \theta(\vec{R})| \quad (3)$$

where w is a weight term controlling the effect of Social Value Orientation on U_i.

In constructing this utility function, we follow a standard approach in mixed-motive reinforcement learning research [24, 28, 56] which provides agents with an overall reward signal combining extrinsic and intrinsic reward [51]. This approach parallels interdependence theory, wherein the effective matrix is formed from the combination of the given matrix and the outcome transformation, and ultimately serves as the basis of actors' decisions [29].

For the exploratory work we detail in the following sections, we restrict our experiments to SVO in the non-negative quadrant (all $\theta \in [0°, 90°]$). The preference profiles in the non-negative quadrant provide the closest match to parameterizations in previous multi-agent research on reward sharing. Nonetheless, we note that interesting preference profiles exist throughout the entire ring [37].

2.3 Algorithm

We deploy advantage actor-critic (A2C) as the learning algorithm for our agents [35]. A2C maintains both value (critic) and policy (actor) estimates using a deep neural network. The policy is updated according to the REINFORCE policy-gradient method, using a value estimate as a baseline to reduce variance. Our neural network
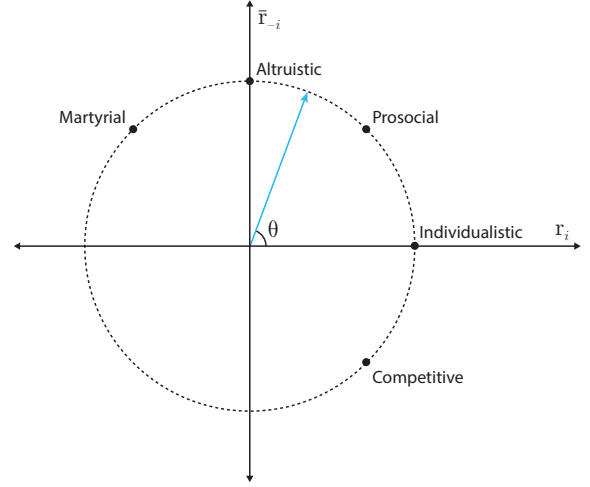

Figure 3: Reward angles and the ring formulation of Social Value Orientation (SVO). Reward angles are a scalar representation of the tradeoff between an agent's own reward and the reward of other agents in the environment. The reward angle an agent prefers is its SVO.

comprises a convolutional layer, a feedforward module, an LSTM with contrastive predictive coding [41], and linear readouts for policy and value. We apply temporal smoothing to observed rewards within the model's intrinsic motivation function, as described by [24].

We use a distributed, asynchronous framework for training [56]. We train populations of $N = 30$ agents with policies $\{\pi_i\}$. For each population, we sample $n = 5$ players at a time to populate each of 100 arenas running in parallel (see also Figure 1a, in which arenas are represented as "samples" from the agent population). Each arena is an instantiation of a single episode of the environment. Within each arena, the sampled agents play an episode of the environment, after which a new group is sampled. Episode trajectories last 1000 steps and are written to queues for learning. Weights are updated from queues using V-Trace [17].

3 MIXED-MOTIVE GAMES

3.1 Intertemporal social dilemmas

For our experiments, we consider two temporally and spatially extended mixed-motive games played with group size $n = 5$: HarvestPatch and Cleanup. These two environments are intertemporal social dilemmas, a particular class of mixed-motive Markov games (c.f., [33]).

Intertemporal social dilemmas are group situations which present a tension between short-term individual incentives and the long-term collective interest [24]. Each individual has the option of behaving prosocially (cooperation) or selfishly (defection). Though unanimous cooperation generates welfare-maximizing outcomes in the long term, on short timescales the personal benefits of acting selfishly strictly dominate those of prosocial behavior. Thus, though all members of the group prefer the rewards of mutual cooperation, the intertemporal incentive structure pushes groups toward
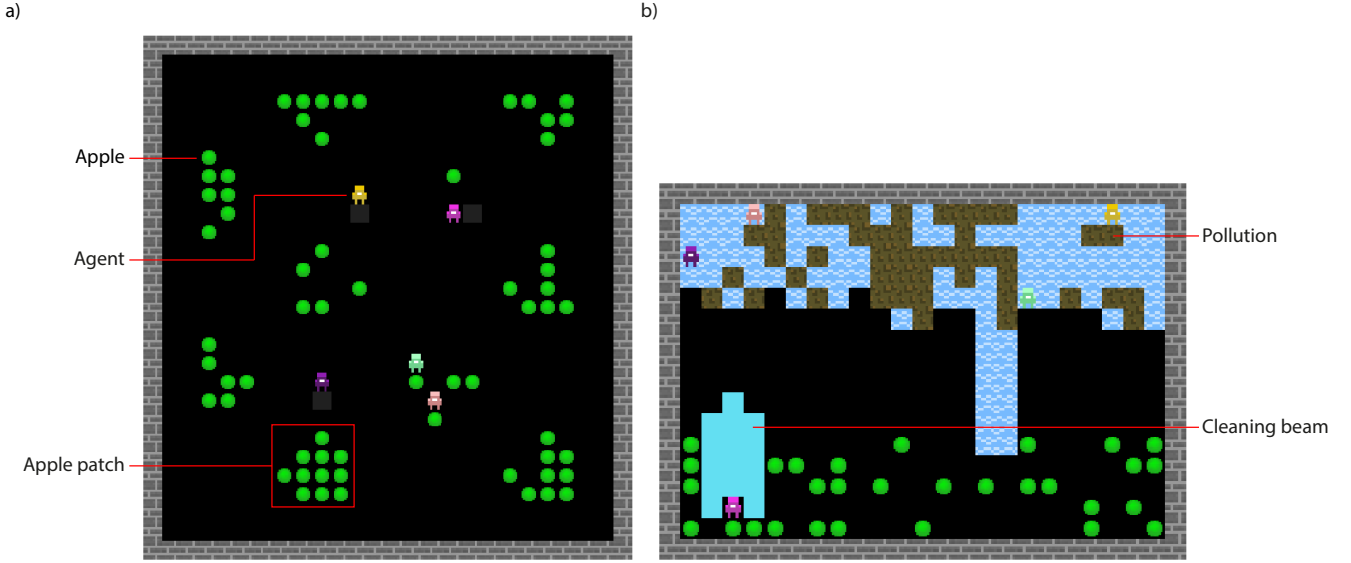
Figure 4: Screenshots of gameplay from (a) HarvestPatch and (b) Cleanup.

welfare-suppressing equilibria. Previous work has evaluated the game theoretic properties of intertemporal social dilemmas [24].

3.2 HarvestPatch

HarvestPatch is a variant of the common-pool resource appropriation game Harvest [24] (Figure 4a). Players are rewarded for collecting apples (reward +1) within a 24 × 26 gridworld environment. Apples regrow after being harvested at a rate dependent on the number of unharvested apples within a regrowth radius of 3. If there are no apples within its radius, an apple cannot regrow. At the beginning of each episode, apples are probabilistically spawned in a hex-like pattern of patches, such that each apple is within the regrowth radius of all other apples in its patch and outside of the regrowth radius of apples in all other patches. This creates localized stock and flow properties [18] for each apple patch. Each patch is irreversibly depleted when all of its apples have been harvested—regardless of how many apples remain in other patches. Players are also able to use a beam to punish other players (reward −50), at a small cost to themselves (reward −1). This enables the possible use of punishment to discourage free-riding [23, 40].

A group can achieve indefinite sustainable harvesting by abstaining from eating "endangered apples" (apples which are the last unharvested apple remaining in their patch). However, the reward for sustainable harvesting only manifests after a period of regrowth if all players abstain. In contrast, an individual is immediately and unilaterally guaranteed the reward for eating an endangered apple if it acts greedily. This creates a dilemma juxtaposing the short-term individual temptation to maximize reward through unsustainable behavior and the long-term group interest of generating higher reward by acting sustainably.

In HarvestPatch, episodes last 1000 steps. Each agent's observability is limited to a 15 × 15 RGB window, centered on its current location. The action space consists of movement, rotation, and use of the punishment beam (8 actions total).

3.3 Cleanup

Cleanup [24] is a public goods game (Figure 4b). Players are again rewarded for collecting apples (reward +1) within a 25 × 18 gridworld environment. In Cleanup, apples grow in an orchard at a rate inversely related to the cleanliness of a nearby river. The river accumulates pollution with a constant probability over time. Beyond a certain threshold of pollution, the apple growth rate in the orchard drops to zero. Players have an additional action allowing them to clean a small amount of pollution from the river. However, the cleaning action only works on pollution within a small distance in front of the agents, requiring them to physically leave the apple orchard to clean the river. Thus, players maintain the public good of orchard regrowth through effortful contributions. As in HarvestPatch, players are able to use a beam to punish other players (reward −50), at a small cost to themselves (reward −1).

A group can achieve continuous apple growth in the orchard by keeping the pollution levels of the river consistently low over time. However, on short timescales, each player would prefer to collect apples in the orchard while other players provide the public good in the river. This creates a tension between the short-term individual incentive to maximize reward by staying in the orchard and the long-term group interest of maintaining the public good through sustained contributions over time.

Episodes last 1000 steps. Agent observability is again limited to a 15 × 15 RGB window, centered on the agent's current location. In Cleanup, agents have an additional action for cleaning (9 actions total).

4 RESULTS

4.1 Social diversity and agent generality

We began by training 12 homogeneous populations per task, with $N = 30$: four consisting of individualistic agents (all $\theta = 0°$), four of prosocial agents (all $\theta = 45°$), and four of altruistic agents (all
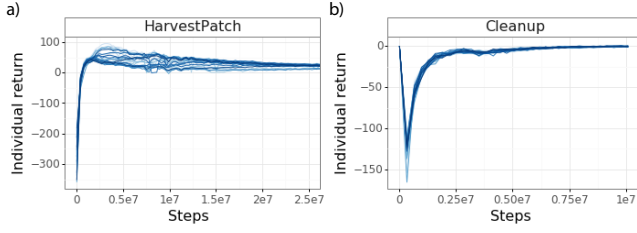
Figure 5: Episode rewards for homogeneous selfish populations playing (a) HarvestPatch and (b) Cleanup. Each individual line shows a single agent's return over training.

	Mean SVO	HarvestPatch	Cleanup
Homogeneous	0°	587.6 (101.7)	-9.9 (11.7)
	45°	665.9 (52.4)	1.1 (2.1)
	90°	**1552.7 (248.2)**	563.8 (235.2)
Heterogeneous	15°	553.4 (574.6)	-0.1 (5.7)
	30°	658.7 (107.1)	2.0 (2.4)
	45°	764.1 (236.3)	6.3 (7.1)
	60°	860.9 (121.5)	318.5 (335.0)
	75°	1167.9 (232.6)	**1938.5 (560.6)**

Table 1: Mean collective returns achieved at equilibrium by homogeneous and heterogeneous populations. Standard deviations are reported in parentheses.

$\theta = 90°$) agents. These resemble previous approaches using selfishness [42], inequity aversion [24, 56], and strong reward sharing [44, 52], respectively. The population training curves for homogeneous selfish populations closely resembled group training curves from previous studies [24, 42] (see sample population training curves in Figure 5). In particular, performance in both environments generated negative returns at the beginning of training due to high-frequency use of the punishment beam. Agents quickly improved performance by learning not to punish one another, but failed to learn cooperative policies. Ultimately, selfish agents were unable to consistently avoid the tragedy of the commons in HarvestPatch or provide public goods in Cleanup.

Optimal hyperparameter values may vary between HarvestPatch and Cleanup. Thus, we selected the weight values w for the two tasks by conducting an initial sweep over w with homogeneous populations of $N = 20$ altruistic agents (all $\theta = 90°$). In HarvestPatch, a weight $w = 0.2$ produced the highest collective returns across several runs (Figure 6a). In Cleanup, a weight $w = 0.1$ produced the highest collective returns across several runs (Figure 6b).

As expected, in HarvestPatch, the highest collective returns among the homogeneous populations were achieved by the altruistic populations (Table 1, Homogeneous row). The prosocial and individualistic populations performed substantially worse. In Cleanup, the highest collective returns similarly emerged among the altruistic populations. The populations of prosocial and individualistic agents, in contrast, achieved near-zero collective returns.

We next trained 80 heterogeneous populations per task. To generate each heterogeneous population, we sampled $N = 30$ SVO values from a normal distribution with a specified mean and dispersion. Since we treated SVO as a bounded variable for these initial experiments, we selected five equally spaced values from 15° to 75° to act as population means and and four equally spaced values from 5.6° ($\frac{\pi}{32}$ radians) to 11.3° ($\frac{\pi}{16}$ radians) to act as population standard deviations. For each mean-standard deviation pair, we generated four populations using different random seeds. We used the same weights w as for the homogeneous populations.

Among the heterogeneous populations, we observed the highest equilibrium collective returns among the 75° population (Table 1, Heterogeneous row). In HarvestPatch, the performance of homogeneous altruistic populations outstripped the performance of the 75° populations. In Cleanup, the reverse pattern emerged: the highest collective returns among all populations are achieved by the heterogeneous 75° populations.

We unexpectedly find that homogeneous populations of altruistic agents produced lower equality scores than most other homogeneous and heterogeneous populations (Table 2). Homogeneous, altruistic populations earned relatively high collective returns in both tasks. However, in each case the produced rewards were concentrated in a small proportion of the population. Agents
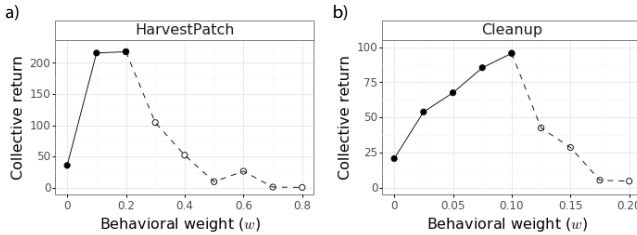

Figure 6: Equilibrium collective return for homogeneous populations of altruistic agents in (a) HarvestPatch and (b) Cleanup. Closed dots reflect populations in which all agents receive positive returns at equilibrium. Open dots indicate populations in which some agents receive zero or negative reward.

	Mean SVO	HarvestPatch	Cleanup
Homogeneous	0°	0.90 (0.09)	0.16 (0.09)
	45°	**0.97 (0.01)**	0.54 (0.07)
	90°	0.29 (0.03)	0.41 (0.08)
Heterogeneous	15°	0.90 (0.13)	0.34 (0.09)
	30°	0.94 (0.06)	0.40 (0.09)
	45°	0.95 (0.03)	0.38 (0.10)
	60°	0.91 (0.02)	0.64 (0.21)
	75°	0.76 (0.04)	**0.87 (0.08)**

Table 2: Mean equality scores achieved at equilibrium by homogeneous and heterogeneous populations. Equality is calculated as the inverse Gini coefficient, with 0 representing all reward being received by a single agent and 1 representing all agents receiving an identical positive reward. Standard deviations are reported in parentheses.
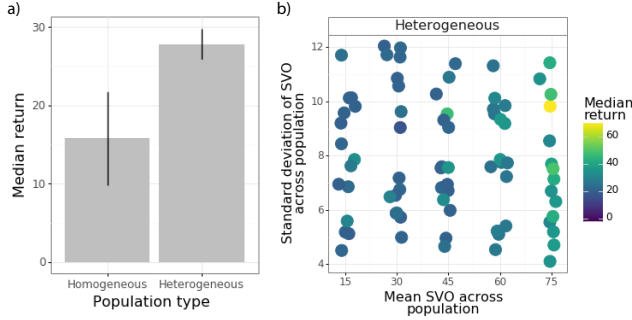
Figure 7: Equilibrium performance of homogeneous and heterogeneous agent populations in HarvestPatch. (a) Heterogeneous populations enjoyed significantly higher median return at equilibrium. (b) Among heterogeneous populations, the highest median returns emerged among the populations whose SVO distributions have both a high mean and a high standard deviation.
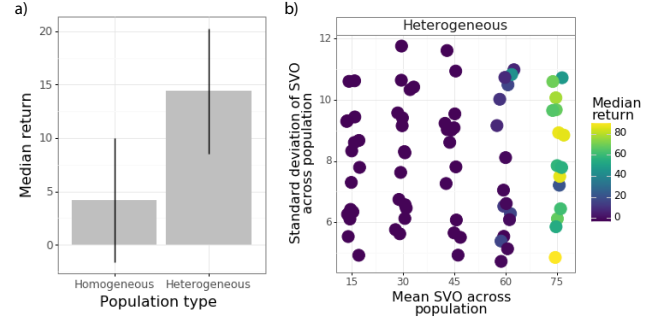

Figure 8: Equilibrium performance of homogeneous and heterogeneous agent populations in Cleanup. (a) Heterogeneous populations received higher median return at equilibrium. (b) Among heterogeneous populations, the highest median returns emerged among the populations with a high mean SVO.

in these homogeneous populations appear to adopt a lazy-agent approach [52] to resolve the conflict between the group's shared preferences for selfless reward distributions. To break the symmetry of this dilemma, most agents in the population selflessly support collective action, thereby optimizing for their social preferences. A smaller number of agents then specialize in accepting the generated reward—shouldering the "burden" of being selfish, in contravention of their intrinsic preferences. This result highlights a drawback of using collective return as a performance metric. Though collective return is the traditional social outcome metric used in multi-agent reinforcement learning, it can mask high levels of inequality.

We therefore revisited population performance by measuring median return, which incorporates signal concerning the efficiency *and* the equality of a group's outcome distribution [8]. Median return can help estimate the generality of learned policies within homogeneous and heterogeneous populations. We compare median return for the two population types by measuring the median return for each population after it reaches equilibrium. We conduct a Welch's *t*-test and report the resulting *t*-statistic, degrees of freedom, and *p*-value. We subsequently provide effect estimates (β) and *p*-values from linear models regressing median return on the mean and standard deviation of population SVO.

Figures 7 and 8 show the generality of policies trained in Harvest-Patch and Cleanup, respectively. In HarvestPatch, heterogeneous populations enjoyed significantly higher median return ($\mu = 27.8$) than homogeneous populations ($\mu = 15.8$) at equilibrium, $t(13.7) = 4.21, p < 0.001$ (Figure 7a). Among heterogeneous populations, a clear pattern could be observed: the higher the population mean SVO, the higher the median return received (Figure 7b). Specifically for populations with high mean SVO, median return appeared to increase slightly when the SVO distribution was more dispersed. When tested with a linear model regressing median return on the mean and standard deviation of SVO, these trends primarily manifested as an interaction effect between mean population SVO and the standard deviation of population SVO, $\beta = 0.025, p = 0.030$. In Cleanup, heterogeneous populations received significantly higher

median return ($\mu = 14.4$) than homogeneous populations ($\mu = 4.2$) at equilibrium, $t(35.9) = 2.43, p = 0.020$ (Figure 8a). Among heterogeneous populations, the highest median returns were observed in tandem with high mean SVO (Figure 8b). However, in this case the effect of the interaction between mean SVO and standard deviation of SVO was non-significant, $p = 0.30$.

In summary, our comparison of homogeneous and heterogeneous populations shows that populations of altruists performed effectively in traditional terms (collective return). However, these populations produced highly specialized agents, resulting in undesirably low equality metrics. Populations with diverse SVO distributions were able to circumvent this symmetry-breaking problem and achieve high levels of median return in HarvestPatch and Cleanup.

4.2 Social preferences and prosocial behavior

How exactly does the distribution of SVO help diverse populations resolve these social dilemmas? We next evaluated the behavioral effects of SVO by examining a single, heterogeneous population within each task. We randomly selected two populations that
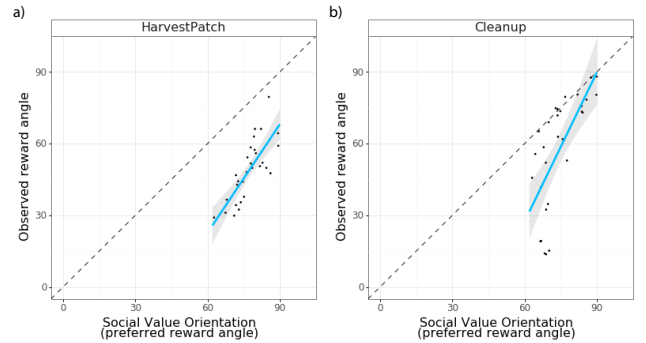

Figure 9: Correlation between target and observed reward angles among SVO agents in (a) HarvestPatch and (b) Cleanup. The higher an agent's SVO, the higher the reward angles it tended to observe.
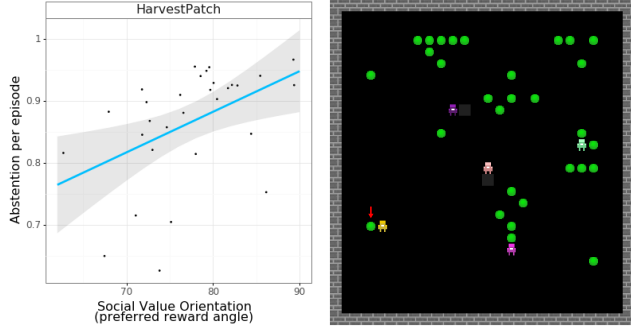
Figure 10: SVO and prosocial behavior in HarvestPatch. Agents with higher SVOs were significantly more likely to abstain from depleting local resource stocks. Here the yellow agent faces the choice of consuming an endangered apple for immediate reward or abstaining and traveling to a different patch.

Figure 11: SVO and prosocial conventions in HarvestPatch. The higher an agent's SVO, the more distance it tended to maintain from other agents in its environment. Here the teal agent is maintaining a particularly high interagent distance, allowing it to sustainably harvest from a single patch.

achieved high equilibrium performance during training, parameterized with mean SVOs of 75° and standard deviations of 7.5°. We gathered data from 100 episodes of play for both of these evaluation experiments, sampling 5 agents randomly for each episode. All regressions reported in this section are mixed error-component models, incorporating a random effect to account for the repeated sampling of individual agents. The accompanying figures depict average values per agent, with superimposed regression lines representing the fixed effect estimate (β) of SVO.

In our evaluation experiments, we observed a positive relationship between an agent's target reward angle and the group reward angles it tended to observe in HarvestPatch, $\beta = 1.51$, $p < 0.001$ (Figure 9a). The effect of SVO on observed reward angle was similarly significant in Cleanup, $\beta = 2.08$, $p < 0.001$ (Figure 9b). This reflects the association of higher agent SVO with the realization of more-prosocial distributions. In both environments, the estimated effect lies below the 45° line, indicating that agents acted somewhat more selfishly than their SVO would suggest.

In HarvestPatch, an agent's prosociality can be estimated by measuring its abstention from consuming endangered apples. We calculated abstention as an episode-level metric incorporating the number of endangered apples an agent consumed and a normalization factor encoding at what points in the episode the endangered apples were consumed. An abstention score of 1 indicates that an agent did not eat a single endangered apple (or that it ate one or more endangered apples on the final step of the episode). An abstention score of 0, though not technically achievable, would indicate that an agent consumed one endangered apple from every apple patch in the environment on the first step of the episode. We observe a significant and positive relationship between an agent's SVO and its abstention, $\beta = 0.0065$, $p = 0.006$ (Figure 10).

The structure of the HarvestPatch environment creates localized stock and flow components. Hosting too many agents in a single patch threatens to quickly deplete the local resource pool. Thus, one rule groups can use to maintain sustainability is for group members to harvest in separate patches, rather than harvesting together and sequentially destroying the environment's apple patches.
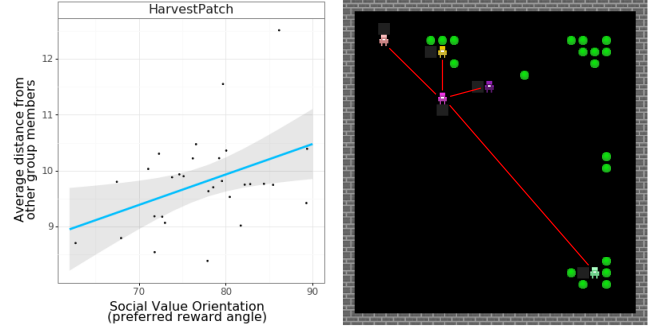
We find that SVO correlated with distance maintained from other group members, $\beta = 0.005$, $p = 0.016$ (Figure 11). Consequently, groups with higher mean SVO established stronger conventions of interagent distance. This simple behavioral convention helped higher-SVO groups guard against environmental collapse.

In Cleanup, an agent's prosociality can be estimated by measuring the amount of pollution it cleans from the river. There was a significant and positive relationship between an agent's SVO and the amount of pollution it cleaned, $\beta = 1.68$, $p = 0.001$ (Figure 12). Agents with higher SVOs acted more prosocially by making greater contributions to the public good.

Finally, do SVO agents develop any sort of prosocial conventions in Cleanup to help maintain high levels of river cleanliness? In Cleanup, we examined one potential coordinating convention that we term behavioral preparedness: an inclination to transition from harvesting to cleaning even before the orchard is fully depleted. In Cleanup, groups that follow the short-term, individual-level incentive structure will respond primarily to the depletion of the
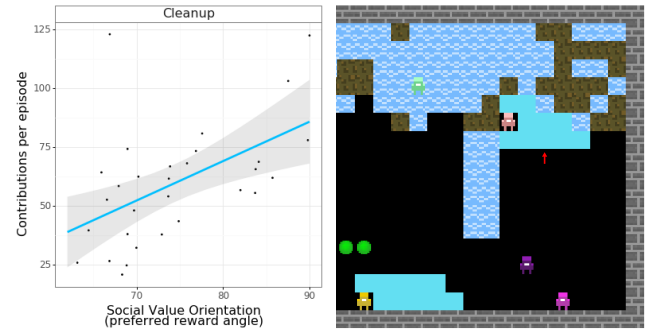

Figure 12: SVO and prosocial behavior in Cleanup. Agents with higher SVOs cleaned a significantly greater amount of pollution per episode than did peers with low SVO. Here the pink agent is actively cleaning two cells of pollution from the river. The yellow agent is using its cleaning action outside of the river, which does not affect its contribution score.
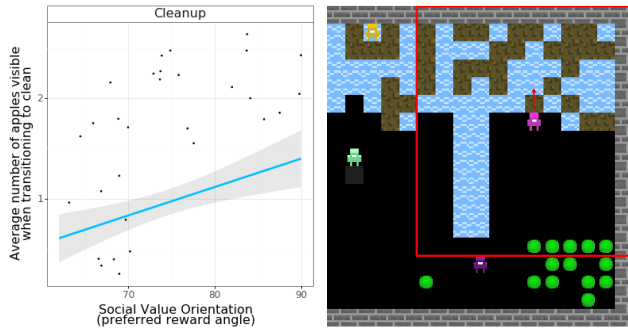
Figure 13: SVO and prosocial conventions in HarvestPatch. Agents with higher SVOs were significantly more likely to enter the river while there were unharvested apples within view. Here the magenta agent is transitioning to clean the river, even though it can observe multiple unharvested apples in the orchard.

orchard, rather than acting preventatively to ensure the public good is sustained over time. Groups that adopt welfare-maximizing strategies, on the other hand, will not wait for the orchard to be fully harvested to clean the river. We find a positive relationship between the average number of apples observable to agents at the times of their transitions to cleaning in each episode, $\beta = 0.028$, $p < 0.001$. The size and significance of this effect are not meaningfully affected by controlling for the number of times each agent transitioned to the river in a given episode, $\beta = 0.028$, $p < 0.001$ (Figure 13). In aggregate, this behavioral pattern helped high-SVO groups maintain higher levels of orchard regrowth over time.

5 DISCUSSION

Recent research on pure-conflict and pure-cooperation reinforcement learning has highlighted the importance of developing robustness to diversity in opponent and partner policies [11, 26, 55]. We extend this argument to the mixed-motive setting, focusing in particular on the effects of heterogeneity in social preferences. Drawing from interdependence theory, we endow agents with Social Value Orientation (SVO), a flexible formulation for reward sharing among group members.

In the mixed-motive games considered here, homogeneous populations of pure altruists achieved high collective returns. However, these populations tended to produce hyper-specialized agents which reaped reward primarily from either intrinsic or extrinsic motivation, rather than both—a method of breaking the symmetry of the shared motivation structure. Thus, when equality-sensitive metrics are considered, populations with diverse distributions of SVO values were able to outperform homogeneous populations.

This pattern echoes the historic observation from interdependence theory that, if both players in a two-player matrix game adopt a "maximize other's outcome" transformation process, the resulting effective matrices often produce deficient group outcomes:

"It must be noted first that, in a number of matrices with a mutual interest in prosocial transformations, if one person acts according to such a transformation, the other is better off by

acting according to his own *given* outcomes than by adopting a similar transformation." [29]

This quote highlights a striking parallel between our findings and the predictions of interdependence theory. We believe this is indicative of a broader overlap in perspective and interests between multi-agent reinforcement learning and the social-behavioral sciences. Here we capitalize on this overlap, drawing inspiration from social psychology to formalize a general mechanism for reward sharing. Moving forward, SVO agents can be leveraged as a modeling tool for social psychology research [36].

In this vein, *group formation* is a topic important to both fields. It is well established among psychologists that an individual's behavior is strongly guided by their ingroup—the group with which they psychologically identify [14]. However, the processes by which individuals form group identities are still being studied and investigated [10, 54]. What sort of mechanisms transform and redefine self-interest to incorporate the interests of a broader group? This line of inquiry has potential linkages to the study of team and coalition formation in multi-agent research [48].

Our findings show that in multi-agent environments, heterogeneous distributions of SVO can generate high levels of population performance. A natural question follows from these results: how can we identify optimal SVO distributions for a given environment? Evolutionary approaches to reinforcement learning [27] could be applied to study the variation in optimal distributions of SVO across individual environments. We note that our results mirror findings from evolutionary biology that across-individual genetic diversity can produce group-wide benefits [39]. We suspect that SVO agents can be leveraged *in simulo* to study open questions concerning the emergence and adaptiveness of human altruism [9, 34].

The development of human-compatible agents still faces major challenges [2, 3, 25]. In pure-common interest reinforcement learning, robustness to partner heterogeneity is seen as an important step toward human compatibility [11]. The same holds true for mixed-motive contexts. Within "hybrid systems" containing humans and artificial agents [12], agents should be able to predict and respond to a range of potential partner behaviors. Social preferences are, of course, an important determinant of human behavior [4, 29]. Endowing agents with SVO is a promising path forward for training diverse agent populations, expanding the capacity of agents to adapt to human behavior, and fostering positive human-agent interdependence.

REFERENCES

[1] Christopher Amato, Girish Chowdhary, Alborz Geramifard, N. Kemal Üre, and Mykel J. Kochenderfer. Decentralized control of partially observable Markov decision processes. In *52nd IEEE Conference on Decision and Control*, pages 2398–2405. IEEE, 2013.

[2] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[4] Daniel Balliet, Craig Parks, and Jeff Joireman. Social Value Orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations*, 12(4):533–547, 2009.

[5] Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748*, 2017.

[6] Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The Hanabi challenge: A new frontier for AI research. *arXiv preprint arXiv:1902.00506*, 2019.

[7] C. Daniel Batson. A history of prosocial research. In A. W. Kruglanski and W. Stroebe, editors, *Handbook of the History of Social Psychology*, pages 243–264. Psychology Press, 2012.

[8] Tony A. Blakely and Ichiro Kawachi. What is the difference between controlling for mean versus median income in analyses of income inequality? *Journal of Epidemiology & Community Health*, 55(5):352–353, 2001.

[9] Samuel Bowles. Group competition, reproductive leveling, and the evolution of human altruism. *Science*, 314(5805):1569–1572, 2006.

[10] Marilynn B. Brewer. In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, 86(2):307, 1979.

[11] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-AI coordination. *arXiv preprint arXiv:1910.05789*, 2019.

[12] Nicholas Christakis. *Blueprint: The Evolutionary Origins of a Good Society*. Hachette, 2019.

[13] David J. Cooper and John H. Kagel. Other-regarding preferences: A selective survey of experimental results. In *The Handbook of Experimental Economics*, volume 2, pages 217–289. Princeton University Press, 2016.

[14] Carsten K. W. de Dreu. Social value orientation moderates ingroup love but not outgroup hate in competitive intergroup conflict. *Group Processes & Intergroup Relations*, 13(6):701–713, 2010.

[15] Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z. Leibo. The Imitation Game: Learned reciprocity in markov games. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1934–1936. International Foundation for Autonomous Agents and Multi-Agent Systems, 2019.

[16] Catherine C. Eckel and Philip J. Grossman. Altruism in anonymous dictator games. *Games and Economic Behavior*, 16(2):181–191, 1996.

[17] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-RL with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

[18] Roy Gardner, Elinor Ostrom, and James M. Walker. The nature of common-pool resource problems. *Rationality and Society*, 2(3):335–358, 1990.

[19] Andrew G. Haldane and Arthur E. Turrell. Drawing on different disciplines: Macroeconomic agent-based models. *Journal of Evolutionary Economics*, 29(1):39–66, 2019.

[20] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

[21] James E. Hartley. Retrospectives: The origins of the representative agent. *Journal of Economic Perspectives*, 10(2):169–177, 1996.

[22] Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

[23] Joseph Henrich. Cooperation, punishment, and the evolution of human institutions. *Science*, 312(5770):60–61, 2006.

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

[25] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, 1(11):517–521, 2019.

[26] Max Jaderberg, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castañeda, Charles Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

[27] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

[28] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, D. J. Strouse, Joel Z. Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pages 3040–3049, 2019.

[29] Harold H. Kelley and John W. Thibaut. *Interpersonal Relations: A Theory of Interdependence*. John Wiley & Sons, 1978.

[30] Alan P. Kirman. Whom or what does the representative individual represent? *Journal of Economic Perspectives*, 6(2):117–136, 1992.

[31] 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 Multi-Agent Systems*, pages 464–473. International Foundation for Autonomous Agents and Multi-Agent Systems, 2017.

[32] Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*, 2017.

[33] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pages 157–163. Elsevier, 1994.

[34] Joshua Mitteldorf and David Sloan Wilson. Population viscosity and the evolution of altruism. *Journal of Theoretical Biology*, 204(4):481–496, 2000.

[35] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

[36] Margaret Morrison and S. Mary Morgan. Models as mediating instruments. In *Models as Mediators*, chapter 2. Cambridge University Press, 1999.

[37] Ryan O. Murphy and Kurt A. Ackermann. Social Value Orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1):13–41, 2014.

[38] Richard E. Nisbett and Ziva Kunda. Perception of social distributions. *Journal of Personality and Social Psychology*, 48(2):297, 1985.

[39] P. Nonacs and K. M. Kapheim. Social heterosis and the maintenance of genetic diversity. *Journal of Evolutionary Biology*, 20(6):2253–2265, 2007.

[40] Rick O'Gorman, Joseph Henrich, and Mark Van Vugt. Constraining free riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 276(1655):323–329, 2008.

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[42] Julien Perolat, Joel Z. Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, pages 3643–3652, 2017.

[43] Alexander Peysakhovich and Adam Lerer. Consequentialist conditional cooperation in social dilemmas with imperfect information. *arXiv preprint arXiv:1710.06975*, 2017.

[44] Alexander Peysakhovich and Adam Lerer. Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 2043–2044. International Foundation for Autonomous Agents and Multi-Agent Systems, 2018.

[45] J. Philippe Rushton, Roland D. Chrisjohn, and G. Cynthia Fekken. The altruistic personality and the self-report altruism scale. *Personality and Individual Differences*, 2(4):293–302, 1981.

[46] Tuomas W. Sandholm and Robert H. Crites. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37(1-2):147–166, 1996.

[47] Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

[48] Prakash P. Shenoy. On coalition formation: A game-theoretical approach. *International Journal of Game Theory*, 8(3):133–164, 1979.

[49] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

[50] Herbert A. Simon. Altruism and economics. *The American Economic Review*, 83(2):156–161, 1993.

[51] Satinder Singh, Andrew G. Barto, and Nuttapong Chentanez. Intrinsically motivated reinforcement learning. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pages 1281–1288. MIT Press, 2004.

[52] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 2085–2087. International Foundation for Autonomous Agents and Multi-Agent Systems, 2018.

[53] Albert W. Tucker. A two-person dilemma. Unpublished, 1950.

[54] John C. Turner. Social categorization and the self-concept: A social cognitive theory of group behavior. In Tom Postmes and Nyla R. Branscombe, editors, *Key Readings in Social Psychology: Rediscovering Social Identity*, pages 243–272. Psychology Press, 2010.

[55] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, pages 1–5, 2019.

[56] Jane X. Wang, Edward Hughes, Chrisantha Fernando, Wojciech M. Czarnecki, Edgar A. Duéñez-Guzmán, and Joel Z. Leibo. Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 683–692. International Foundation for Autonomous Agents and Multi-Agent Systems, 2019.