Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents

Elizaveta Tennant[1], **Stephen Hailes**[1], **Mirco Musolesi**[1, 2]

[1]Department of Computer Science
University College London, Gower Street, London WC1E 6BT, United Kingdom
[2]Department of Computer Science and Engineering
University of Bologna, Viale del Risorgimento 2, 40133 Bologna, Italy
l.karmannaya.16@ucl.ac.uk, s.hailes@ucl.ac.uk, m.musolesi@ucl.ac.uk

Abstract

Growing concerns about safety and alignment of AI systems highlight the importance of embedding *moral* capabilities in artificial agents: a promising solution is the use of learning from experience, i.e., Reinforcement Learning. In multi-agent (social) environments, complex population-level phenomena may emerge from interactions between individual learning agents. Many of the existing studies rely on simulated social dilemma environments to study the interactions of independent learning agents; however, they tend to ignore the moral *heterogeneity* that is likely to be present in societies of agents in practice. For example, at different points in time a single learning agent may face opponents who are consequentialist (i.e., focused on maximizing outcomes over time), norm-based (i.e., conforming to specific norms), or virtue-based (i.e., considering a combination of different virtues). The extent to which agents' co-development may be impacted by such moral heterogeneity in populations is not well understood. In this paper, we present a study of the learning dynamics of morally heterogeneous populations interacting in a social dilemma setting. Using an Iterated Prisoner's Dilemma environment with a partner selection mechanism, we investigate the extent to which the prevalence of diverse moral agents in populations affects individual agents' learning behaviors and emergent population-level outcomes. We observe several types of non-trivial interactions between pro-social and anti-social agents, and find that certain types of moral agents are able to steer selfish agents towards more cooperative behavior.

Introduction

Growing concerns about the safety and alignment of Artificial Intelligence (AI) systems highlight the importance of embedding pro-social capabilities into artificial agents. Without these, agents might potentially behave harmfully in multi-agent situations such as social dilemmas (e.g., see Leibo et al. 2017). In particular, in this work, we focus on modeling the development of *moral decision-making*.

In general, it has been shown that morality can be developed in agents through learning from experience (Hughes et al. 2018; McKee et al. 2020; Tennant, Hailes, and Musolesi 2023b). Learning approaches to moral alignment offer a variety of advantages, including potentially greater adaptability and generality (Tennant, Hailes, and Musolesi 2023a), and the ability to learn implicit preferences (Ouyang et al. 2022). Moral principles can range from consequentialist (Bentham 1780) to norm-based morality (Kant 1785), and from entirely pro-social to entirely anti-social preferences. Furthermore, the morality of agents can be based on a combination of virtues, as exemplified by the Virtue Ethics approach (Aristotle n.a.). Existing works have proposed a variety of individual models to predispose agents to specific moral principles (e.g., Bazzan, Bordini, and Campbell 1999; Capraro and Perc 2021), and recent studies have demonstrated that these can be effectively encoded in Reinforcement Learning (RL) agents via intrinsic rewards (Hughes et al. 2018; McKee et al. 2020; Tennant, Hailes, and Musolesi 2023b).

However, to date, we still have limited understanding of how *different* types of morality may co-evolve in heterogeneous populations. In fact, many real-world AI systems are likely to co-exist (essentially forming systems of systems) and may be co-developed in parallel with others. Especially when it comes to integrating moral decision-making into systems, different stakeholders may decide to prioritize varying principles or preferences. It is therefore crucial that we start developing an understanding of how the presence of different moral preferences in populations affects individual agents' learning behaviors and interactions, and, indeed, emergent population-level outcomes. Furthermore, simulation studies such as ours contribute to raising awareness about emergent behaviors and the possibility of unintuitive outcomes emerging from multi-agent learning scenarios.

This type of agent-based analysis of moral learning may additionally offer insights into the potential dynamics of interactions in human societies in the tradition of computational philosophy (McKenzie 2007), provide an experimental test-bed for new theories (Mayo-Wilson and Zollman 2021), or simulate potential evolutionary mechanisms underpinning human moral preferences, in a similar way to Evolutionary Game Theory (Hofbauer and Sigmund 1998; Sigmund and Nowak 1999; Sigmund 2010).

The core contribution of this work is the *study of behavior and population dynamics among RL agents with diverse moral preferences*, providing insights for the design of ar-

Version with Appendix: https://arxiv.org/abs/2403.04202
Code: https://github.com/Liza-Tennant/moral_MARL

tificial agents with a focus on safety and alignment. More generally, we present a *methodology for analyzing emergent behavior in populations of agents with heterogeneous moral orientation*. We apply the proposed methodology to the Iterated Prisoner's Dilemma (IPD, Rapoport 1974), demonstrating at the same time its potential applicability and generalizability to a variety of scenarios and different moral frameworks.

Background and Preliminaries

Social Dilemma Games

Social dilemma games simulate social situations in which players obtain different utilities (payoffs) from choosing one action or another, and the structure of these utilities is such that each player faces a trade-off between individual interest and societal benefit when choosing an action (Dawes 1980). The most widely studied type is a symmetric matrix game with two players and two abstract actions - Cooperate (C) or Defect (D). Players in these games must decide on their respective actions simultaneously, without communicating.

A classic game from Economics and Philosophy, which is relevant to moral choice, is the Prisoner's Dilemma (Rapoport 1974, see payoffs for row vs column player in Table 1). We implement the *Iterated* Prisoner's Dilemma (IPD) in a population, in which agents interact in pairs in discrete time steps and aim to maximize their cumulative payoff over time. In the iterated version of the game, players can take actions to punish their opponents for past defection or to influence their future behaviors.

IPD	C	D
C	3, 3	0, 4
D	4, 0	1, 1

Table 1: Payoffs on one step of the *Iterated Prisoner's Dilemma (IPD)* game for the row & column player.

Deriving predicted equilibria for these games is not always feasible: these are systems composed of interacting (heterogeneous) entities, where learning dynamics may cause instability in the environment and dynamic behaviors (Busoniu, Babuska, and De Schutter 2008). Therefore, simulation is needed to study potential emergent behaviors and outcomes (Anderson 1972).

Deep Reinforcement Learning in Markov Games

Reinforcement Learning (RL) is a well-suited technique for modeling agents that learn by interacting with others in an environment (Sutton and Barto 2018). It can be applied in conjunction with Evolutionary Game Theory (Hofbauer and Sigmund 1998) to iterated social dilemma games (Littman 1994; Sandholm and Crites 1996; de Cote, Lazaric, and Restelli 2006; Abel, MacGlashan, and Littman 2016), in which game payoffs constitute *extrinsic* rewards, and traits such as moral or social preferences (Fehr and Fischbacher 2002) can be encoded in the agent's *intrinsic* reward (Chentanez, Barto, and Singh 2004). In the following section, we discuss the design of intrinsic moral rewards in detail. It is also worth noting that we assume populations of independent, continuously learning agents. This creates interesting dynamics as agents affect one another's learning process (Leibo et al. 2019). Given the potentially large number of states, we adopt DQN as the underlying learning algorithm (Mnih et al. 2015).

Morality as Intrinsic Reward

Traditional social dilemma scenarios assume that agents are only motivated by accumulating game payoffs (i.e., extrinsic reward). However, human data has shown that other preferences may also come into play, such as the predisposition to cooperation (Camerer 2011). In artificial agents, preferences other than game rewards can be encoded in intrinsic rewards (Chentanez, Barto, and Singh 2004). As a result, recent multi-agent RL studies have been focused on modeling various pro-social preferences as intrinsic reward functions for social dilemma players (Hughes et al. 2018; Peysakhovich and Lerer 2018; McKee et al. 2020). Moral rewards in particular have been studied for two-agent (i.e., dyadic) interactions in Tennant, Hailes, and Musolesi (2023b).

We anchor our intrinsic reward definitions in traditional moral philosophical frameworks, especially relying on the distinction between *consequentialist* versus *norm-based* morality, and the idea of *virtue ethics*. *Consequentialist* morality focuses on the consequences of an action, and includes Utilitarianism (Bentham 1780), which defines actions as moral if they maximize total utility for all agents in a society. *Norm*-based morality, including Deontological ethics (Kant 1785), considers an act moral if it does not contradict the society's external norms. Finally, in *Virtue Ethics* (Aristotle n.a.), moral agents must act in line with their certain internal virtues, such as fairness or care for others (Graham et al. 2013). Different virtues can matter more or less to different agents (Aristotle n.a.; Graham, Haidt, and Nosek 2009) and can themselves have consequentialist or norm-based foundations. We present the specific set of agents considered in this study and their classification in the Methodology section.

Partner Selection

In human populations, agents have a choice of which individual to interact with. Given this selection mechanism, reputation comes into play and competitive and collaborative relationships may form. Santos, Santos, and Pacheco (2008) show experimentally that these dynamics are likely to lead to a re-structuring of the population. Adaptive behavior resulting from selection mechanisms has been hypothesized to drive the emergence of cooperation (Barclay and Willer 2007; Cuesta et al. 2015). In the context of societies of artificial learning agents, these mechanisms have recently been studied by Anastassacos, Hailes, and Musolesi (2020) and Baker (2020).

In particular, when applying partner selection to the IPD, a conflict may arise between a player's motivation to appear cooperative in order to get selected more often by other payers, and the motivation to select and then exploit cooperators to gain a greater payoff. Anastassacos, Hailes, and Musolesi (2020) find partner selection to be norm-inducing and

even lead to the emergence of cooperation in a population of purely selfish agents. We adopt a partner selection model similar to Anastassacos, Hailes, and Musolesi (2020).

Selection dynamics in morally heterogeneous populations are likely to be more complex and harder to predict. Since each player plays according to their own intrinsic reward signal, different coalitions (Shenoy 1979) may arise among different agent types within a population, and popularity of agents may not directly correlate with cooperativeness. Running simulation experiments to study this may provide insight into the types of behaviors and outcomes that may develop in heterogeneous populations with selection, with implications not only for the study of artificial agents, but also of human and animal societies and Evolutionary Game Theory (Sigmund and Nowak 1999).

Learning in Heterogeneous Populations

Evidence from the social sciences suggests that human societies are morally heterogeneous (Graham et al. 2013; Bentahila, Fontaine, and Pennequin 2021). While certain commonsense norms may be agreed upon by a society (Reid and Haakonssen 1990; Gert 2004), distinct moral principles are likely to hold a different weight for different individuals depending on factors such as political orientation (Graham, Haidt, and Nosek 2009), personality (Lifton 1985), nurture or even natural predispositions (Sinnott-Armstrong 2008).

McKee et al. (2020) conducted a social dilemma study with diverse consequentialist agents (i.e., agents with various preferences over group outcome distributions). Their findings suggest that agents trained in heterogeneous populations develop particularly generalized and high-performing policies. In contrast, populations containing only altruistic agents are characterized by greater collective reward, but at the cost of equality between agents. In addition, in McKee et al. (2022), the authors define population diversity as the set of various opponents' policies that an agent may face, and also find that training in diverse populations can lead to improvements in terms of agent performance on some types of environments.

Even without learning, Santos, Santos, and Pacheco (2008) show that - in fixed networks of agents - diversity in terms of neighborhoods within the population graph promotes cooperation in the population. With respect to global societal outcomes, Ord (2015) discusses how, through interaction, morally diverse agents taking actions to satisfy their distinct moral goals may improve the global welfare of the whole population, akin to a 'trade'. We believe that these conceptual frameworks can represent the basis for studying how artificial agents may learn to act across populations containing different proportions of opponents with different moral preferences.

Methodology

Experimental Setup

Our environment involves a population P (of size $N = 16$) playing an iterated game. At each iteration, a moral player M and an opponent O play a simultaneous one-shot Prisoner's Dilemma game with two possible actions - Cooperate or Defect ($a_M^t, a_O^t \in \{C, D\}$, see Table 1 for the respective payoffs). To allow learning in a population, we run our simulation in 30000 episodes, where a single episode involves each player selecting a partner once, and then each corresponding pair of players playing the game. For clarity, each agent gets to make a selection on each episode, so N games are played. In particular, at iteration t, the learning agent M observes a selection state including the action played by each possible opponent $O \in P$ at $t-1$: $s_{M,sel}^t = [a_{O_1}^{t-1}, ..., a_{O_{n-1}}^{t-1}]$. Using this state and the current learned policy, M selects an opponent O. Then, to choose an action for the dilemma, player M relies on the state representing the latest move of their selected opponent O: $s_{M,dil}^t = a_O^{t-1}$. Simultaneously, their opponent O chooses an action following the same principle. The players M and O then each receive a game reward R_M^{t+1} and R_O^{t+1} (corresponding to the game payoffs) and observe a new state $s_{M,dil}^{t+1}$ and $s_{O,dil}^{t+1}$ based on their opponent's move a^t.

In this Markov game (Littman 1994), over time both players learn to make selections and take actions by observing past interactions with various opponents. Each agent uses Q-Learning (Watkins and Dayan 1992) to update the state-action value function Q (i.e., an estimate of the cumulative expected return over time). Given a large number of possible selection states, we approximate Q using a neural network. Each agent maintains two internal models (one for partner selection, and one for playing the dilemma) - for consistency, we use function approximation for learning both models, and each is parameterized by a fully connected network with a single hidden layer of size 256.

For learning both selections and the dilemma, players record the experiences (s, a, r, s') in their memory buffer D_{sel} and D_{dil}. We copy the reward obtained from the game into both the dilemma and selection memories. At the end of an episode, all players simultaneously update the Q-network parameters θ_t for their two models using the latest experience available in the memory buffer D, and the Mean Squared Error loss function (Mnih et al. 2015), where $\gamma = 0.99$ is a discount factor:

$$L_t(\theta_t) = E_{s,a,r,s'}\left[\left(R^{t+1} + \gamma \max_a Q(s^{t+1}, a, \theta_t) - Q(s^t, a^t, \theta_t)\right)^2\right] \quad (1)$$

If more than one experience is available (i.e., a player plays more than one dilemma game in that episode), we calculate an average loss across the experiences before updating the network. Each memory buffer D is refreshed before the start of the next episode, so that each player only learns from the latest episode of experience. In updating the network weights with the Adam optimizer (Kingma and Ba 2015), we use a learning rate of 0.001.

Agents select partners and play the dilemma using an ϵ-greedy policy, acting randomly with probability ϵ or otherwise playing greedily according to the Q-values learned so far:

$$\pi(s_t) = \begin{cases} \arg\max_a Q(s_t, a), & \text{with probability } 1-\epsilon \\ U(A = \{C, D\}), & \text{with probability } \epsilon. \end{cases} \quad (2)$$

	Agent Label & Moral Type		*Moral Reward Function*	*Source of Morality*
	S	Selfish	None (use $R^t_{M_{extr}}$ to learn)	None
Pro-Social	*Ut*	Utilitarian	$R^t_{M_{intr}} = R^t_{M_{extr}} + R^t_{O_{extr}}$	External/Internal consequentialist
	De	Deontological	$R^t_{M_{intr}} = \begin{cases} -\xi, & \text{if } a^t_M = D, a^{t-1}_O = C \\ 0, & \text{otherwise} \end{cases}$	External norm
	V-Eq	Virtue-Equality	$R^t_{M_{intr}} = 1 - \frac{\lvert R^t_{M_{extr}} - R^t_{O_{extr}} \rvert}{R^t_{M_{extr}} + R^t_{O_{extr}}}$	Internal consequentialist
	V-Ki	Virtue-Kindness	$R^t_{M_{intr}} = \begin{cases} \xi, & \text{if } a^t_M = C \\ 0, & \text{otherwise} \end{cases}$	Internal norm
Anti-Social	*aUt*	Anti-Utilitarian	$R^t_{M_{intr}} = -(R^t_{M_{extr}} + R^t_{O_{extr}})$	External/Internal consequentialist
	mDe	Malicious Deontological	$R^t_{M_{intr}} = \begin{cases} \xi, & \text{if } a^t_M = D, a^{t-1}_O = C \\ 0, & \text{otherwise} \end{cases}$	External norm
	V-In	Virtue-Inequality	$R^t_{M_{intr}} = \frac{\lvert R^t_{M_{extr}} - R^t_{O_{extr}} \rvert}{R^t_{M_{extr}} + R^t_{O_{extr}}}$	Internal consequentialist
	V-Ag	Virtue-Aggression	$R^t_{M_{intr}} = \begin{cases} \xi, & \text{if } a^t_M = D \\ 0, & \text{otherwise} \end{cases}$	Internal norm

Table 2: Definitions of the types of intrinsic moral rewards, from the point of view of the moral agent M playing versus an opponent O at iteration t. We define four pro-social players, four anti-social ones, and the traditional *Selfish* player.

We set $\epsilon_{sel} = 0.1$ and $\epsilon_{dil} = 0.05$. The full detailed algorithm can be found in the Appendix.

Player M can learn according to an *extrinsic* game reward $R^{t+1}_{M_{extr}}$, which depends directly on the joint actions a^t_M, a^t_O (as defined in Table 1), or according to an *intrinsic* moral reward $R^{t+1}_{M_{intr}}$. In the next subsection, we discuss the definition of these intrinsic moral rewards.

Modeling Morality as Intrinsic Reward

Intrinsic rewards associated with pro-social preferences have been used to incentivize the emergence of cooperation in social dilemmas (Hughes et al. 2018; Peysakhovich and Lerer 2018; Jaques et al. 2019). Tennant, Hailes, and Musolesi (2023b) extended this work by defining four Q-learning moral[1] agents that learn according to various intrinsic rewards $R_{M_{intr}}$, and contrasting these against a traditional *Selfish* agent which learns to maximize its extrinsic (game) reward $R_{M_{extr}}$ (e.g., Leibo et al. 2017).

In this study, we rely on eight types of moral agents - four pro-social moral learners from Tennant, Hailes, and Musolesi (2023b) *(Utilitarian, Deontological, Virtue-Equality, Virtue-Kindness)*, and four additional anti-social counterparts *(anti-Utilitarian, malicious-Deontological, Virtue-Inequality, Virtue-Aggression)*. The agents are defined as follows (for formal definitions of the reward functions, please refer to Table 2):

- the *Utilitarian (Ut)* agent tries to maximize the collective payoff (i.e., total payoff for both players; Bentham 1780);
- the *Deontological (De)* agent tries to follow the norm of conditional cooperation (Kant 1785; Fehr and Fischbacher 2004) and gets penalized through the negative reward $-\xi$ for defecting against a cooperator (i.e., an opponent who previously cooperated);
- the *Virtue-Equality (V-Eq)* agent tries to maximize equality between the two players' payoffs $R^t_{M_{extr}}$ and $R^t_{O_{extr}}$, measured using a two-agent variation of the Gini coefficient (Gini 1912);
- the *Virtue-Kindness (V-Ki)* agent receives a reward ξ for acting kindly (i.e., cooperating) against any opponent (Aristotle n.a.);
- the *anti-Utilitarian (aUt)* agent tries to minimize the collective payoff;
- the *malicious-Deontological (mDe)* agent tries to follow the conditional defection norm and receives a reward ξ for defecting against an cooperator;
- the *Virtue-Inequality (V-In)* agent tries to minimize equality between the two players' payoffs $R^t_{M_{extr}}$ and $R^t_{O_{extr}}$;
- the *Virtue-Aggression (V-Ag)* agent receives a reward ξ for acting aggressively (i.e., defecting) against any opponent.

Agent types *Ut* and *aUt* can be defined as external consequentialist since their reward depends on consequences in the environment (i.e., the players' rewards). The *De* and *mDe* agents depend on an external reputation-based norm defined in terms of current actions. *V-Eq* and *V-In* agents can be considered internal consequentialist, since they follow a consequence-based internal virtue. Finally, *V-Ki* and *V-Ag*'s pre-dispositions originate from the agent's internal norm.

The rewards in Table 2 are defined for a single iteration t. Given the payoff matrices of the IPD game used (see Table 1), we set the parameter ξ in the four norm-based rewards to be $\xi = 5$, so it sends a strong signal of a value similar

[1]We use the term moral for agents that are not selfish. However, selfishness itself can be considered as a moral choice, expressed as rational egotism. In the case of social dilemmas, selfishness maps to the concept of rationality. For these agents $R_{M_{intr}} = R_{M_{extr}}$.

Population Label	*Population Composition*
majority-S	8x*S*, 1x*Ut*, 1x*aUt*, 1x*De*, 1x*mDe*, 1x*V-Eq*, 1x*V-In*, 1x*V-Ki*, 1x*V-Ag*
majority-Ut	1x*S*, 8x*Ut*, 1x*aUt*, 1x*De*, 1x*mDe*, 1x*V-Eq*, 1x*V-In*, 1x*V-Ki*, 1x*V-Ag*
majority-aUt	1x*S*, 1x*Ut*, 8x*aUt*, 1x*De*, 1x*mDe*, 1x*V-Eq*, 1x*V-In*, 1x*V-Ki*, 1x*V-Ag*
majority-De	1x*S*, 1x*Ut*, 1x*aUt*, 8x*De*, 1x*mDe*, 1x*V-Eq*, 1x*V-In*, 1x*V-Ki*, 1x*V-Ag*
majority-mDe	1x*S*, 1x*Ut*, 1x*aUt*, 1x*De*, 8x*mDe*, 1x*V-Eq*, 1x*V-In*, 1x*V-Ki*, 1x*V-Ag*
majority-V-Eq	1x*S*, 1x*Ut*, 1x*aUt*, 1x*De*, 1x*mDe*, 8x*V-Eq*, 1x*V-In*, 1x*V-Ki*, 1x*V-Ag*
majority-V-In	1x*S*, 1x*Ut*, 1x*aUt*, 1x*De*, 1x*mDe*, 1x*V-Eq*, 8x*V-In*, 1x*V-Ki*, 1x*V-Ag*
majority-V-Ki	1x*S*, 1x*Ut*, 1x*aUt*, 1x*De*, 1x*mDe*, 1x*V-Eq*, 1x*V-In*, 8x*V-Ki*, 1x*V-Ag*
majority-V-Ag	1x*S*, 1x*Ut*, 1x*aUt*, 1x*De*, 1x*mDe*, 1x*V-Eq*, 1x*V-In*, 1x*V-Ki*, 8x*V-Ag*

Table 3: Populations considered in this study. Each population contains eight players of a certain 'majority' type and one player of each other type.

to the maximum game payoff available. Through further experiments, we also observe that the choice of smaller values of ξ does not affect the overall results.

Population Compositions

We compare a set of heterogeneous populations based on the following principles: population size is equal to 16; all populations contain at least one player of each type; one player type constitutes the majority (i.e., 8 out of 16 players). Given 9 possible player types, this results in 9 possible population compositions, as outlined in Table 3.

Results

We separately analyze cooperation (at population and individual level), social outcomes (at population level), selection dynamics and rewards obtained by different player types across the nine populations.

Emergence of Cooperation

Cooperation across the entire population. We first study whether a stronger prevalence of certain agent types in every population leads to higher levels of cooperation in the population as a whole. Cooperation over time for each population is presented in Figure 1a. Results show that the greatest level of cooperation is achieved by the *majority-Ut* and *majority-V-Ki* populations, with around 70% of the players cooperating by the end. In the *majority-V-Ki* population, this high level of cooperation develops early on but then drops to about 60% around episode 9000, which is then followed by an increase back to 70% soon after, and a stabilization around episode 13000. An analysis of behavior by each player type (see Appendix, Figures 7a & 7b) shows that *V-Ki* players in particular go through an initial period of increasing defection - this coincides with an instability in the whole population where defection is still preferable because it increases the probability of being selected by others. Eventually, many of the players learn to avoid defectors and the *V-Ki* players learn their optimal policy of nearly-full cooperation.

In the *majority-Ut* population, cooperation emerges more slowly, remaining around 60% for the first 10000 episodes, but then stabilizes without further drops after episode 13000 (at a similar point in time as the *majority-V-Ki* population).

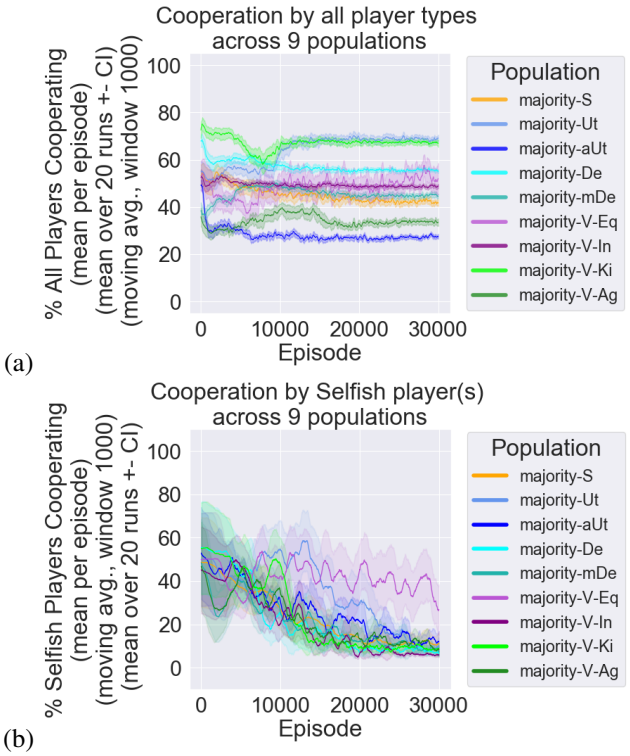

Figure 1: **(a)** Cooperation by all player types within each population over time. **(b)** Cooperation by the *Selfish* player(s) in every population over time. In these charts, we plot the moving average of the mean across 20 runs.

Specifically, *Ut* players take longer than *V-Ki* to learn to cooperate (see Appendix, Figures 7a & 7b), but once cooperation levels increase to around 70%, it remains stable.

The other majority-prosocial populations *majority-De* and *majority-V-Eq* do not promote as much cooperation. In the *majority-De* population, cooperation drops quickly in the first 1000 episodes and then stabilizes at 60% - lower than for *majority-Ut* and *majority-V-Ki* populations. An analysis of action pairs in each population (see Appendix, Figure 6) shows that *De* players in this population learn an near-100% cooperative strategy very early on. The low level of cooperation in the *majority-De* population can therefore be ex-

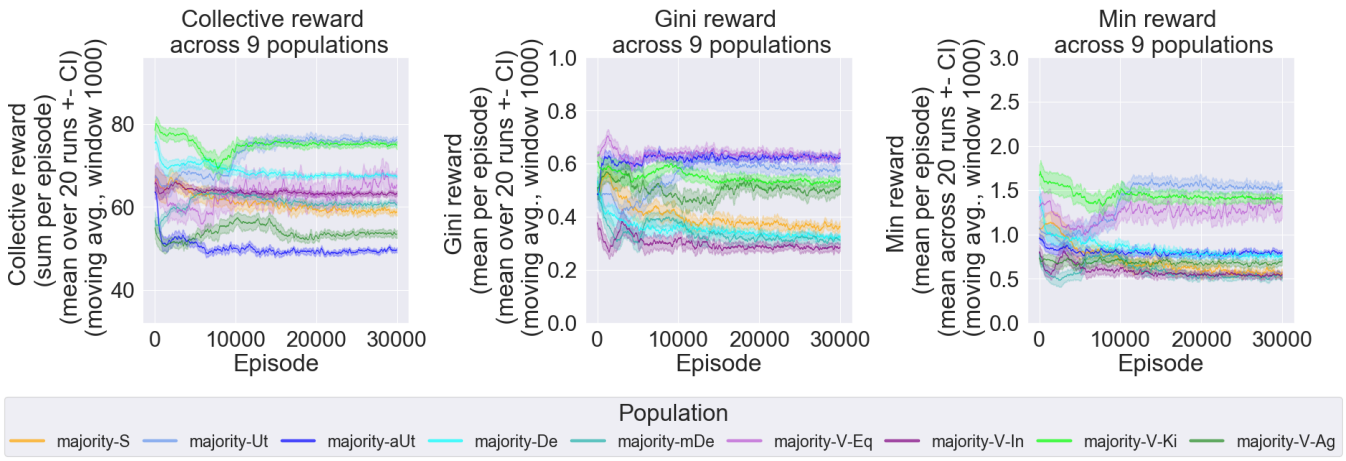

Figure 2: Population-level social outcomes over time: Collective Reward, Gini Reward and Min Reward. We plot the moving average of the mean across 20 runs.

plained by the behavioral dynamics involving other player types: this population in particular is characterized by high levels of exploitation (see Appendix, Figure 6), whereby the *De* players select aggressive opponents such as *aUt* or *V-Ag*, and those opponents consistently defect against the *De* players, thus bringing the overall level of cooperation in the population down. We discuss the corresponding selection dynamics later in this section.

Finally, both the *majority-V-Eq* and the *majority-V-In* populations achieve similar levels of cooperation, which are lower than for any of the other majority-prosocial population. This can be explained by the fact that both the *V-Eq* and *V-In* players never learn fully cooperative policies (see Appendix, Figure 7b) - due to the structure of the IPD game, where both (C,C) and (D,D) actions lead to equal outcomes. The difference between these populations is that *majority-V-Eq* displays much more instability in terms of cooperation per episode (see Figure 1), and more mutual cooperation or defection than the *majority-V-In* population (see Appendix, Figure 6).

The least amount of cooperation is observed in the *majority-aUt* population. An analysis of pairs of actions over time (see Appendix, Figure 6) shows that this is due to the prevalence of mutual defection between any pair of players in particular, and follows strategically from the fact that mutual defection allows for the lowest collective payoff in the IPD, satisfying the goals of the *aUt* agents.

Behavior of selfish learners. In addition, we analyze the cooperative behavior of *S* learners in each population. These players are the most aligned to traditional multi-agent learning literature, since they do not learn according to any intrinsic reward, but rather by simply maximizing the game reward according to the IPD payoff matrix. This analysis investigates which population compositions steer *S* learners into more cooperative behaviors. An analysis of cooperative behavior exhibited by each non-*S* player type across populations is available in Appendix, Figures 7a & 7b.

As shown in Figure 1b, *S* players generally learn a mostly-defective policy in most populations, though cooperation is slightly above the 5% chance level even at the end of the simulation, likely due to pressure to cooperate to get selected by some other players. *S* players display most cooperation in the *majority-V-Eq* and the *majority-Ut* populations, though in the latter cooperation still decreases to 10% towards the second half of the training period. Thus, the *majority-V-Eq* population has a unique influence on the learning of the *S* agent, and is able to steer this agent towards a greater level of cooperation for longer. The *V-Eq* opponent is about 40% likely to cooperate in this population (see Appendix, Figure 7b), which must drive the *S* player to also learn the near-40% cooperative policy observed here.

Social Outcomes

In addition to cooperation levels, we analyze a set of social outcomes for each population. We adopt standard social outcome metrics used in the existing literature (Hughes et al. 2018; Tennant, Hailes, and Musolesi 2023b) - in particular, we measure collective reward ($R_{collective}$), equality between payoffs (R_{gini}) and the lowest payoff obtained (R_{min}) for a pair of agents $\{M, O\}$ that play on any iteration t. To aggregate values from every iteration t within an episode (where the length of the episode is equal to the population size $N = 16$), we sum $R_{collective}$, and average R_{gini} and R_{min} per episode as follows:

$$R_{collective} = \sum_{t=1}^{N} (R^t_{M_{extr}} + R^t_{O_{extr}}) \tag{3}$$

$$R_{gini} = \frac{\sum_{t=1}^{N} (1 - \frac{|R^t_{M_{extr}} - R^t_{O_{extr}}|}{R^t_{M_{extr}} + R^t_{O_{extr}}})}{N} \tag{4}$$

$$R_{min} = \frac{\sum_{t=1}^{N} \min(R^t_{M_{extr}}, R^t_{O_{extr}})}{N}. \tag{5}$$

In Figure 2 we present the outcomes over time in each population. Collective reward follows a similar pattern to

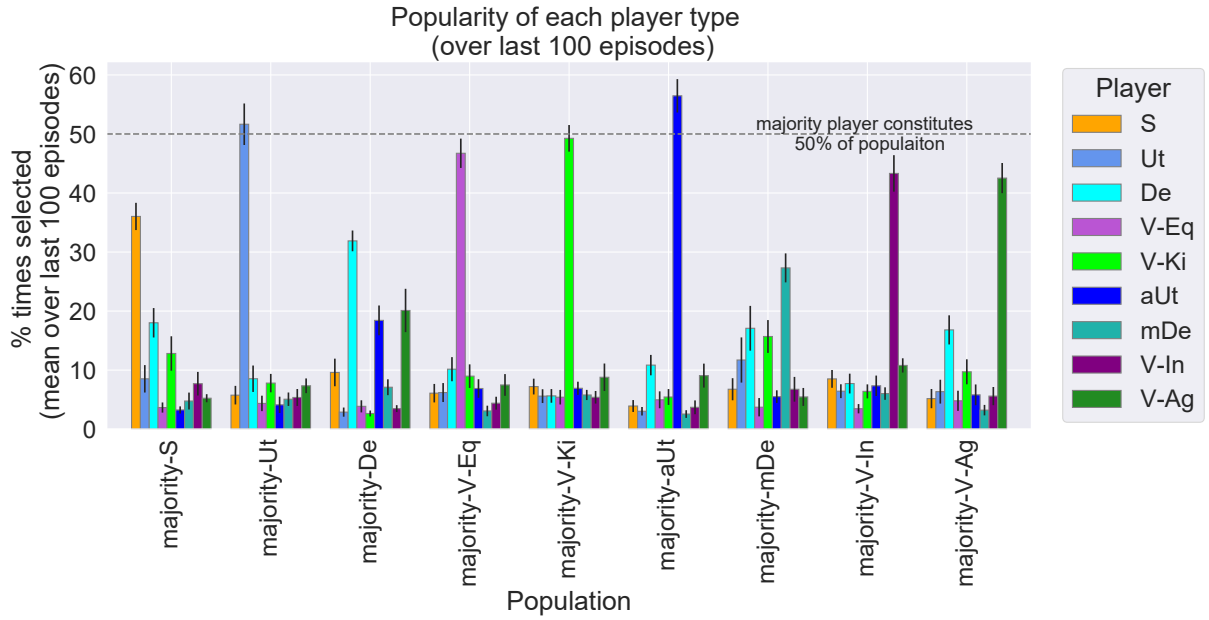

Figure 3: Popularity of player types in each population on the final 100 episodes. Values represent the average across 20 runs and the associated confidence intervals. We sum the values for cases where more than one player of the same type is present (e.g. 8x*S* players in the *majority-S* population). For ease of interpretation, we add a 50% reference line - this allows us to compare whether the majority player is selected more (or less) often than expected simply due to their prevalence in each population.

population-level cooperation (Figure 1a), since these are interconnected in the IPD (in particular, cooperation by one or both of the players on the IPD leads to greater collective reward than mutual defection, as shown by the payoffs in Table 1).

Equality between the players (i.e., Gini reward) is highest for the equality-focused *majority-V-Eq* population - though the absolute mean value per episode does not go above 0.7, meaning that unequal outcomes remain prevalent even in this population. Interestingly, an analysis of action pairs (see Appendix, Figure 6) shows that this level of equality is achieved in this population equally through mutual cooperation and mutual defection. Equality is also generally high for the pro-social populations *majority-Ut* and *majority-V-Ki* due to the prevalence of mutual cooperation, and for their anti-social counterparts *majority-aUt* and *majority-V-Ag* due to the prevalence of mutual defection. Equality is lower when the majority of players are of the *S, De, mDe* or *V-Ie* types, since in these scenarios one player tends to exploit the other.

Minimum reward is highest once again in the *majority-Ut* and the *majority-V-Ki* populations (with values around 1.5, due to the prevalence of mutual cooperation - see Appendix, Figure 6), and also in the *majority-V-Eq* case (due to a lack of exploitation). All other populations result in lower minimum reward of values between 0.5 and 1.0, which are explained by either exploitation of mutual defection in the IPD.

Selection Dynamics

Given the selection mechanism implemented, it is interesting to analyze which players are the most popular overall in each population (Figure 3), and which selector types prefer which types of partners (Figure 4). In a population of purely selfish agents (Anastassacos, Hailes, and Musolesi 2020), the selection mechanism adds an incentive to cooperate, since selfish players benefit from playing against cooperators and tend to select them. In our heterogeneous populations, however, certain players may not necessarily prefer cooperative opponents, so selection dynamics are likely to be more complex.

Figure 3 shows the popularity of each player type in each population over the final 100 episodes. Generally, in probabilistic terms, especially at the beginning, the majority player type tends to be the most popular in their respective population. However, we observe that in certain populations, for example *majority-S* and *majority-De*, the majority player type is selected less than 50% of the time - i.e., much less frequently than expected due to their prevalence. In these cases, it is interesting to analyze which non-majority players emerge as popular alternatives. For the full analysis of selection patterns across individual players in all populations, see Appendix, Figures 8a, 8b, 9a, 9b.

The behavior of the *majority-De* population is the most striking. Here, *aUt* and *V-Ag* players are relatively popular. A further analysis of the specific selection patterns among individual players in this population (Figure 4b) reveals that the *De* players in particular tend to mostly prefer anti-social opponents *aUt* and *V-Ag*. This apparently self-sabotaging

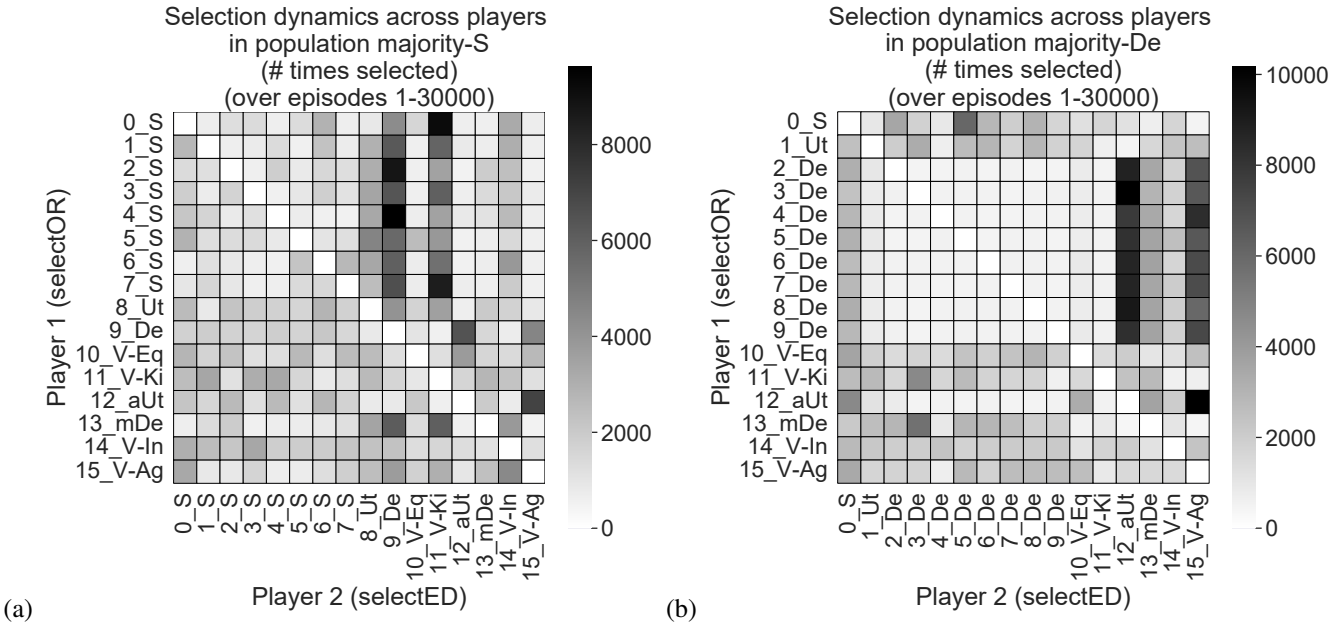

Figure 4: Selections made by individual players, with number of selections summed over all 30000 episodes (average across 20 runs), for two example populations: *majority-S* **(a)** & *majority-De* **(b)**; for all populations, see Appendix, Figures 8a, 8b, 9a, 9b.

behavior can be explained by a closer analysis of the *De* player's R_{intr} - they are penalized for defecting against a cooperator, so the safest behavior in a heterogeneous population is to always select a defector, so that they have the lowest chance of violating their moral norm. Figure 4b also shows that another player - *aUt* - prefers to avoid the *De* majority players and instead tends to select the *V-Ag* opponent. This highlights that minimizing collective reward (which is the preference embedded in the *aUt* agent's intrinsic reward) is best achieved by avoiding the very cooperative *De* agent. Thus, we find an interesting dynamic in the heterogeneous *majority-De* population, in which the majority of the players follow a strong cooperative norm, demonstrating ways in which these *De* players can be very cooperative (see Figure 1a and Appendix, Figure 6) but not very popular.

Returning to the analysis of player popularity on the final 100 episodes (Figure 3), the *majority-mDe* population also has a few non-majority players emerging as popular. In this case, it is the *De* and *V-Ki* players that get selected often by others. Thus, in both *majority-De* and *majority-mDe* populations, the norm-following majority player is not even popular among their own kind.

Finally, the *majority-S* population is also interesting to analyze (Figure 4a), because this population differs from others in that most agents here are pure payoff-maximizers in terms of the traditional IPD game. One might expect that *S* players would often select *De* or *V-Ki* opponents, since these are so easily exploitable, but in fact *Ut* and *V-In* also emerge as relatively popular, likely because these players will have learned to cooperate against a defector, which benefits the *S* player in terms of reward. An analysis of action pairs (see Appendix, Figure 6) confirms that unilateral defection is indeed common in this population. Many of the other players in this population tend to select majority type *S* opponents roughly as often as other player types, meaning that *Selfish* players do not emerge as unpopular, despite their often defective behavior. The only exceptions to this are *mDe* and *V-Ag*. The *mDe* player in particular learns to avoid the *S* opponent since *mDe* would rather face a cooperator (such as *De* or *V-Ki*), and obtain positive R_{intr}, than select the *S* players which are likely to defect.

Cumulative Reward

Finally, we analyze the cumulative reward obtained for each individual player type in each population. Figure 5 presents the reward accumulated by each player type in each population over all 30000 episodes. The results in terms of game reward clearly show that - in general - pro-social players are disadvantaged on the IPD game itself, since they get exploited and/or do not benefit from exploiting a cooperative opponent. Furthermore, the lowest-performing player type is *De*. However, the *De* player obtains high levels of intrinsic reward in many scenarios, especially when their own kind constitute the majority of the population. Other pro-social players obtain the highest R_{intr} in populations in which some anti-social player type constitutes the majority.

Anti-social or *S* players obtain higher game reward over time than their pro-social counterparts - likely by exploiting others. The *majority-De* population allows *aUt* and *V-Ag* players to obtain especially high levels of game reward; given the selection dynamics analyzed above, we understand that this is driven by the fact that *De* players actively prefer to select these opponents and then get exploited by them. The lowest-performing players in terms of intrinsic reward in general are the norm-based anti-social players *mDe* and *V-Ag* (especially in majority-anti-social popula-

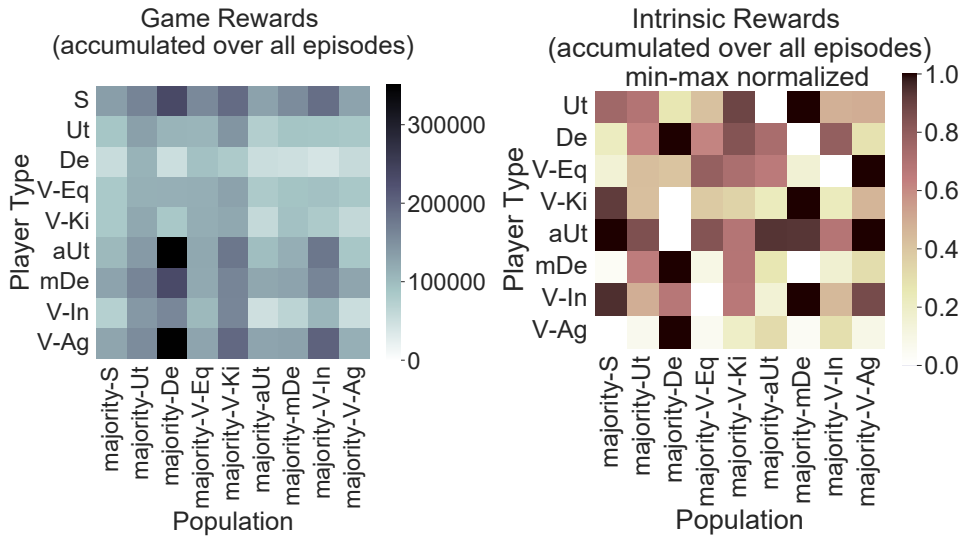

Figure 5: Game reward and intrinsic reward accumulated by each player type in each population over the entire 30000 episodes (averaged across 20 runs). We average per number of players of a certain type (e.g. 8x*S* players in the *majority-S* population). For comparability, we normalize intrinsic rewards using the minimum and maximum observed value for each player type.

tions) and the norm-based prosocial players *majority-V-Ki* (especially in majority-pro-social populations). Players of the *aUt* type obtain high intrinsic reward in most populations except *majority-De* (because the presence of such cooperative players makes it hard to minimize collective reward), whereas the *mDe* and *V-Ag* obtain the highest intrinsic reward by exploiting the very cooperative *De* players in the *majority-De* population.

Discussion

In this paper we have investigated how the presence of different moral preferences in populations affects individual agents' learning behaviors and emergent social outcomes. The results provide insights into behaviors emerging from multi-agent interactions between morally heterogeneous agents. Some of our findings point to the dangers of designing artificial agents in certain ways - for example, showing how this may lead to the development of self-sabotaging behaviors. Additionally, while our study can be considered as a stark simplification of reality, the insights deriving from it can provide a starting point towards investigating the emergence of similar behaviors in human societies. In summary, many of the agents' actions are consistent with their reward definitions - in general, pro-social agents prefer to cooperate (except for the *V-Eq* agent), and anti-social agents display lower levels cooperation and obtain lower collective reward (except for the *V-In* agent).

In terms of temporal dynamics, our results demonstrate interesting asymmetries in the emergence of cooperation by agents with different moral preferences. In particular, we observe that consequentialist agents *Ut* take longer to learn to cooperate than the norm-based agents *De*, while the norm-based agents *V-Ki* go through an unstable and defective period before learning a stable cooperative policy. Furthermore, for certain consequentialist agents (specifically, *aUt* vs *Ut*), the convergence to an equilibrium in the Prisoner's Dilemma environment is faster for anti-social players than their pro-social counterparts - likely due to the payoffs attributed to defection in this game.

We have also observed some surprising interactions between different types of morality due to the selection mechanism present. First, the introduction of a large number of equality-focused agents *(V-Eq)* has a positive effect on the cooperative behavior of a *Selfish* agent *(S)*, providing insight into how diverse opponent types can steer self-interested agents towards more pro-social behavior. Second, one pro-social player type (specifically, the norm-based player *De*) learns to select anti-social opponents in order to avoid violating their own moral norm. Many anti-social players simultaneously learn to select the *De* player and exploit them in the game, since this exploitative behavior is not penalized by anyone in the population and does not deter the *De* player from further interaction. Thus, our results show that the introduction of *De* type agents risks promoting anti-social actors and inequality in a population. This illustrates the possibility that the presence of narrowly-defined norms might lead to self-sabotaging behavior and cause negative outcomes for the population as a whole.

Conclusion

This study is the first to analyze the dynamics of learning in populations of agents with moral preferences varying from those with consequentialist foundations to those with norm-based ones. Our results demonstrate the potential of using intrinsic rewards for modeling various moral preferences in RL agents. More generally, we have provided a generic methodology for studying the learning dynamics of heterogeneous populations, and introduced measures for as-

sessing individual and societal outcomes. Finally, this work might also contribute to raising awareness about emergent behaviors and the possibility of unintuitive outcomes in such multi-agent learning scenarios.

Our research agenda includes the study of more complex moral frameworks and an exploration of multi-objective approaches that combine moral and self-interested motivation. We also plan to investigate environments characterized by learning dynamics under partial observability.

Acknowledgments

Elizaveta Tennant was supported by the Leverhulme Trust (DS-2017-026; Doctoral Training Programme for the Ecological Study of the Brain).

References

Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement learning as a framework for ethical decision making. In *Proceedings of the AAAI Workshop: AI, Ethics, and Society (AIES'16)*, 54–61.

Anastassacos, N.; Hailes, S.; and Musolesi, M. 2020. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*, volume 34, 7047–7054.

Anderson, P. W. 1972. More Is Different. *Science*, 177(4047): 393–396.

Aristotle. n.a. *The Nicomachean Ethics*. Oxford University Press.

Baker, B. 2020. Emergent Reciprocity and Team Formation from Randomized Uncertain Social Preferences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS'20)*, volume 33, 15786–15799.

Barclay, P.; and Willer, R. 2007. Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, 274(1610): 749–753.

Bazzan, A. L. C.; Bordini, R. H.; and Campbell, J. A. 1999. Moral Sentiments in Multi-agent Systems. In Müller, J. P.; Rao, A. S.; and Singh, M. P., eds., *Intelligent Agents V: Agents Theories, Architectures, and Languages*, 113–131. Springer Berlin Heidelberg.

Bentahila, L.; Fontaine, R.; and Pennequin, V. 2021. Universality and cultural diversity in moral reasoning and judgment. *Frontiers in Psychology*, 12.

Bentham, J. 1780. *An Introduction to the Principles of Morals and Legislation.* Clarendon Press.

Busoniu, L.; Babuska, R.; and De Schutter, B. 2008. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2): 156–172.

Camerer, C. F. 2011. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Capraro, V.; and Perc, M. 2021. Mathematical foundations of moral preferences. *Journal of The Royal Society Interface*, 18(175): 20200880.

Chentanez, N.; Barto, A.; and Singh, S. 2004. Intrinsically Motivated Reinforcement Learning. In *Proceedings of the 18th International Conference on Neural Information Processing Systems (NeurIPS'04)*, volume 17, 1281–1288.

Cuesta, J. A.; Gracia-Lázaro, C.; Ferrer, A.; Moreno, Y.; and Sánchez, A. 2015. Reputation drives cooperative behaviour and network formation in human groups. *Scientific Reports*, 5(1): 7843.

Dawes, R. M. 1980. Social Dilemmas. *Annual Review of Psychology*, 31(1): 169–193.

de Cote, E. M.; Lazaric, A.; and Restelli, M. 2006. Learning to cooperate in multi-agent social dilemmas. In *Proceedings of the 5th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'06)*, 783–785.

Fehr, E.; and Fischbacher, U. 2002. Why social preferences matter – the impact of non-selfish motives on competition, cooperation and incentives. *The Economic Journal*, 112(478): 1–33.

Fehr, E.; and Fischbacher, U. 2004. Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4): 185–190.

Gert, B. 2004. *Common Morality: Deciding What to Do*. New York: Oxford University Press.

Gini, C. 1912. *Variabilità e Mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statistiche. [Fasc. I.]*. Tipografia di P. Cuppini.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral Foundations Theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, volume 47, 55–130. Academic Press.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5): 1029–1046.

Hofbauer, J.; and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Hughes, E.; Leibo, J. Z.; Phillips, M.; Tuyls, K.; Dueñez-Guzman, E.; García Castañeda, A.; Dunning, I.; Zhu, T.; McKee, K.; Koster, R.; Zhu, T.; Roff, H.; and Graepel, T. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS'18)*, volume 31, 3330–3340.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*, 3040–3049. PMLR.

Kant, I. 1785. *Grounding for the Metaphysics of Morals*. Cambridge University Press.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*.

Leibo, J. Z.; Hughes, E.; Lanctot, M.; and Graepel, T. 2019. Autocurricula and the Emergence of Innovation

from Social Interaction: A Manifesto for Multi-Agent Intelligence Research. arXiv preprint arXiv:1903.00742. arXiv:1903.00742.

Leibo, J. Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'17)*, 464–473.

Lifton, P. D. 1985. Individual differences in moral development: The relation of sex, gender, and personality to morality. *Journal of Personality*, 53(2): 306–334.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on International Conference on Machine Learning (ICML'94)*, 157–163.

Mayo-Wilson, C.; and Zollman, K. J. S. 2021. The computational philosophy: simulation as a core philosophical method. *Synthese*, 199(1-2): 3647–3673.

McKee, K. R.; Gemp, I.; McWilliams, B.; Duèñez Guzmán, E. A.; Hughes, E.; and Leibo, J. Z. 2020. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'20)*, 869–877.

McKee, K. R.; Leibo, J. Z.; Beattie, C.; and Everett, R. 2022. Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 36(1): 21.

McKenzie, A. J. 2007. *The Structural Evolution of Morality*. Cambridge University Press.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Ord, T. 2015. Moral Trade. *Ethics*, 126(1): 118–138.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS'22)*, volume 35, 27730–27744.

Peysakhovich, A.; and Lerer, A. 2018. Consequentialist conditional cooperation in social dilemmas with imperfect information. In *Proceedings of the 6th International Conference on Learning Representations (ICLR'18)*.

Rapoport, A. 1974. Prisoner's dilemma — recollections and observations. In *Game Theory as a Theory of a Conflict Resolution*, 17–34. Springer.

Reid, T.; and Haakonssen, K. 1990. *Thomas Reid on Practical Ethics*. Princeton University Press.

Sandholm, T. W.; and Crites, R. H. 1996. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37(1-2): 147–166.

Santos, F. C.; Santos, M. D.; and Pacheco, J. M. 2008. Social diversity promotes the emergence of cooperation in public goods games. *Nature*, 454(7201): 213–216.

Shenoy, P. P. 1979. On Coalition Formation: A Game-Theoretical Approach. In *International Journal of Game Theory*, volume 8.

Sigmund, K. 2010. *The Calculus of Selfishness*. Princeton University Press.

Sigmund, K.; and Nowak, M. A. 1999. Evolutionary game theory. *Current Biology*, 9(14): R503–R505.

Sinnott-Armstrong, W. 2008. *Moral Psychology: The Evolution of Morality: Adaptations and Innateness*. MIT Press.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT Press. ISBN 0262039249.

Tennant, E.; Hailes, S.; and Musolesi, M. 2023a. Learning Machine Morality through Experience and Interaction. arXiv Preprint. arXiv:2312.01818.

Tennant, E.; Hailes, S.; and Musolesi, M. 2023b. Modeling Moral Choices in Social Dilemmas with Multi-Agent Reinforcement Learning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI'23)*, 317–325.

Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine Learning*, 8(3): 279–292.