Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)

Modeling Moral Choices in Social Dilemmas with Multi-Agent Reinforcement Learning

Elizaveta Tennant[1] , **Stephen Hailes**[1] , **Mirco Musolesi**[1,2]
[1]University College London
[2]University of Bologna
{l.karmannaya.16, s.hailes, m.musolesi}@ucl.ac.uk

Abstract

Practical uses of Artificial Intelligence (AI) in the real world have demonstrated the importance of embedding *moral choices* into intelligent agents. They have also highlighted that defining top-down ethical constraints on AI according to any one type of morality is extremely challenging and can pose risks. A bottom-up learning approach may be more appropriate for studying and developing ethical behavior in AI agents. In particular, we believe that an interesting and insightful starting point is the analysis of emergent behavior of Reinforcement Learning (RL) agents that act according to a predefined set of *moral rewards* in social dilemmas.

In this work, we present a systematic analysis of the choices made by intrinsically-motivated RL agents whose rewards are based on moral theories. We aim to design reward structures that are simplified yet representative of a set of key ethical systems. Therefore, we first define moral reward functions that distinguish between consequence- and norm-based agents, between morality based on societal norms or internal virtues, and between single- and mixed-virtue (e.g., multi-objective) methodologies. Then, we evaluate our approach by modeling repeated dyadic interactions between learning moral agents in three iterated social dilemma games (Prisoner's Dilemma, Volunteer's Dilemma and Stag Hunt). We analyze the impact of different types of morality on the emergence of cooperation, defection or exploitation, and the corresponding social outcomes. Finally, we discuss the implications of these findings for the development of moral agents in artificial and mixed human-AI societies.

1 Introduction

An open question in AI research and development is how to represent and specify ethical choices and constraints for this class of technologies in computational terms [Ajmeri *et al.*, 2020; Amodei *et al.*, 2016; Awad *et al.*, 2022; Dignum, 2017; Yu *et al.*, 2018; Wallach, 2010]. In particular, there is an increasing interest in understanding how certain types of behavior and outcomes might emerge from the interactions of learning agents in artificial societies [de Cote *et al.*, 2006; Foerster *et al.*, 2018; Hughes *et al.*, 2018; Jaques *et al.*, 2019; Leibo *et al.*, 2017; McKee *et al.*, 2020; Peysakhovich and Lerer, 2018a; Peysakhovich and Lerer, 2018b; Rodriguez-Soto *et al.*, 2021; Sandholm and Crites, 1996] and in interactive systems where humans are in the loop [Carroll *et al.*, 2019; Rahwan *et al.*, 2019]. We believe that a promising and insightful starting point is the analysis of emergent behavior of Reinforcement Learning (RL) agents that act according to a predefined set of *moral rewards* in situations were there is tension between individual interest and collective social outcomes, namely in social dilemmas [Axelrod and Hamilton, 1981; Rapoport, 1974; Sigmund, 2010].

In this work, we present for the first time a systematic analysis of the learning dynamics and choices of RL agents whose rewards are based on moral theories. The goal is to define reward structures that are simplified yet representative of a set of key ethical frameworks. In other words, the (moral) agents play the social dilemmas not trying to maximize their cumulative return considering the payoff matrix of the game, but using their intrinsic reward according to a given ethical system. Classic moral theories include Utilitarianism[1] [Bentham, 1996], Deontological morality [Kant, 1981] and Virtue Ethics [Aristotle, 2009] - each of these explains different aspects of human moral preferences, and has specific implications for social learning agents [Anderson *et al.*, 2005]. We map choices to an intrinsic reward system [Chentanez *et al.*, 2004] according to these ethical frameworks. The majority of existing modeling work has focused on single types of social preference [Hughes *et al.*, 2018; Jaques *et al.*, 2019; Kleiman-Weiner *et al.*, 2017; Peysakhovich and Lerer, 2018b; Peysakhovich and Lerer, 2018a]. However, as suggested by the continued debate between the three moral frameworks [Mabille and Stoker, 2021], and by evidence from human moral psychology [Bentahila *et al.*, 2021; Graham *et al.*, 2009], a broad range of moral preferences are

Version with Appendix: https://arxiv.org/abs/2301.08491
Code: https://github.com/Liza-Tennant/moral_choice_dyadic

[1]In this paper, the term Utilitarian indicates an agent that shows prosocial behavior according to several exponents of Utilitarianism [Bentham, 1996], i.e., one that tries to maximize the global reward given by the sum of the individual rewards, in contrast with a Selfish agent, whose goal is the maximization of their own utility.

likely to exist within and across societies (especially in preferences regarding AI morality [Awad *et al.*, 2018]). Thus, implementing any one theory *top-down* without consideration of how it might interact with other types of moral reasoning risks creating societies in which exploitative and/or defenseless behavior emerges [Wallach and Allen, 2009]. The present study instead is based on a *bottom-up* modeling of interactions between different types of moral learning agents as a way of gathering insights for understanding and improving human, artificial and hybrid societies.

Social dilemma games, which illustrate the trade-off between self-interested and mutually beneficial choices [Rapoport, 1974], create a useful testing ground for moral agents [Cavagnetto and Gahir, 2014; Cunningham, 1967; Hegde *et al.*, 2020]. In particular, in this work we consider *iterated* games. The use of learning agents in these games allows us to study how certain behaviors might evolve in a society given a set of initial agent preferences and environmental characteristics such as payoff structure [Harper *et al.*, 2017; Sandholm and Crites, 1996]. Evolutionary modeling can effectively be applied to agents' moral choices to study the emergence of specific social outcomes [Binmore, 2005]. For example, a growing body of research has explored emergent cooperation within this framework [Jaques *et al.*, 2019; Leibo *et al.*, 2017], with some studies specifically implementing certain aspects of morality to elicit cooperative behaviors [Hughes *et al.*, 2018; Peysakhovich and Lerer, 2018a]. One study [McKee *et al.*, 2020] has moved towards exploring populations heterogeneous in terms of social preferences. However, to our knowledge no work has yet modeled the behavior of different moral agents learning against one another in such social dilemma settings [Tolmeijer *et al.*, 2021], so little is known about the potential risks that may evolve from implementing any one type of morality in AI agents who act in diverse social environments.

The contributions of this work can be summarized as follows:

- We introduce a methodological framework for developing moral artificial agents and define intrinsic moral rewards inspired by philosophical theories.
- We systematically evaluate emergent behaviors and outcomes in pairwise interactions, including between different types of moral agents, in three social dilemma games, namely Iterated Prisoner's Dilemma, Iterated Volunteer's Dilemma and Iterated Stag Hunt.

We believe that our approach can be generalized to other types of moral agents or games, and can be used in the future to model agent learning against human opponents. Also for this reason, the code used for this study is available as open source software to encourage further work in this area.

2 Background and Related Work

2.1 Social Dilemma Games

Social dilemma games simulate social situations in which agents obtain different payoffs from choosing one action or another, and due to the structure of the game, each agent faces a trade-off between individual interest and societal benefit.

IPD	C	D	**IVD**	C	D	**ISH**	C	D
C	3,3	1,4	C	4,4	2,5	C	5,5	1,4
D	4,1	2,2	D	5,2	1,1	D	4,1	2,2

Table 1: Payoff matrices for each step of the Iterated Prisoner's Dilemma (IPD), Iterated Volunteer's Dilemma (IVD) & Iterated Stag Hunt (ISH) games, in which players are motivated to defect by either *greed* (IVD), *fear* (ISH), or both (IPD).

The most widely studied type is a symmetric matrix game with two players, each choosing one of two possible abstract actions - Cooperate or Defect. Players must decide on their respective actions simultaneously, without communicating.

Three classic games from economics and philosophy which are relevant to moral choice are the Prisoner's Dilemma [Rapoport, 1974], Volunteer's Dilemma (or 'Chicken') [Poundstone, 1993], and Stag Hunt [Skyrms, 2001] (see payoffs for row vs. column player in Table 1). We implement *iterated* games, in which players repeatedly face one of these dilemmas, and over numerous interactions learn to maximize their cumulative payoff.

In the Prisoner's Dilemma, mutual cooperation would achieve a Pareto-optimal outcome (in which one player cannot be made better off without disadvantaging the other) - but each individual player's best response is to defect due to *greed* (facing a cooperator, they benefit from defecting) and *fear* (facing a defector, they suffer by cooperating). In the Volunteer's Dilemma, a selfish or rational player may choose to defect due to greed, and if both do so, both obtain the lowest possible payoffs (i.e. no one volunteers, and the society suffers). Finally, in the Stag Hunt game two players can cooperate in hunting a stag and thus obtain the greatest possible payoff each; however, given a lack of trust between the players (i.e., fear of a non-cooperative partner), either may be tempted to defect and hunt a hare on their own instead, decreasing both players' payoffs as a result.

2.2 Repeated Games and Learning

Iterated versions of the matrix games provide a complex set of strategies, since players can take actions to punish their opponents for past wrongdoing or to influence their future behaviors, causing instability in the environment and dynamic behaviors. Thus, calculating predicted equilibria in these situations is not always computationally plausible, and simulation methods are required in order to study potential emergent behaviors and outcomes.

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 [Abel *et al.*, 2016; Littman, 1994; Sandholm and Crites, 1996], in which 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 *et al.*, 2004].

2.3 Moral Choices

The tension between individual and social benefit presents an interesting test-bed for ethics [Cavagnetto and Gahir, 2014]:

how would different moral theories manifest themselves in these three games? Certain embedded social preferences have been shown to promote cooperation between RL agents in social dilemmas - such as learning via opponent-learning awareness [Foerster *et al.*, 2018], social influence [Jaques *et al.*, 2019], or prosocial reward functions [Peysakhovich and Lerer, 2018a], including inequity-aversion [Hughes *et al.*, 2018] and social value orientation [McKee *et al.*, 2020].

We anchor our definitions of moral choices in traditional moral philosophy. Of the numerous moral theories which have been proposed over the millennia, we will focus on three established and influential moral frameworks: consequentialist, norm-based and virtue ethics. The methodological approach presented in this work can potentially be applied to a variety of other moral systems. *Consequentialist* morality focuses on the consequences of an action, and includes Utilitarianism [Bentham, 1996], which defines actions as moral if they maximize total utility for all agents in a society. *Norm-* or *Duty*-based morality, including Deontological ethics [Kant, 1981], considers an act moral if it does not contradict the society's external norms. Finally, in *virtue ethics* [Aristotle, 2009], moral agents must act in line with their certain internal virtues, such as fairness or care [Graham *et al.*, 2013]. Different virtues can matter more or less to different agents [Aristotle, 2009; Graham *et al.*, 2009] and can themselves have consequentialist or norm-based foundations. Furthermore, a single agent may rely on more than one type of virtue, so a more expressive way of modeling virtue ethics might be through a multi-objective paradigm. We will investigate this by implementing a mixed virtue agent alongside single-virtue agents.

In this work we study the implications of each of these theories in social dilemma environments with learning agents. We model moral preferences inspired by these theories though intrinsic reward functions [Chentanez *et al.*, 2004]. Intrinsic rewards have been used to model consequentialist traits, including preferences against inequity [Hughes *et al.*, 2018], towards prosocial [Peysakhovich and Lerer, 2018a; Peysakhovich and Lerer, 2018b] and/or competitive outcomes [Abel *et al.*, 2016; McKee *et al.*, 2020]. Work on norm-based moral agents is limited [Tolmeijer *et al.*, 2021], mainly focused on computational modeling of norm emergence among agents with static policies (i.e., without learning) [Salahshour, 2022], or using learning with reputation mechanisms [Anastassacos *et al.*, 2021]. However, no studies investigated the impact of pre-defined moral norms in social dilemmas, which is the focus of the present work.

It is worth noting that no work has focused on the learning interactions between consequentialist and norm-based moral agents. In the broader social preferences literature, [McKee *et al.*, 2020] has shown that RL agents trained in populations with diverse social preferences learn more robust and general strategies, while prosocial agents in [Foerster *et al.*, 2018; Hughes *et al.*, 2018] got exploited when facing agent types other than themselves. Heterogeneous moralities are observed in human societies [Graham *et al.*, 2013], so as a starting point we explicitly investigate the implications of different moral agent types learning against one another in dyadic interactions. This also gives us an opportunity to test potential criticisms of each moral framework, as well as their interactions, through simulation. In particular, philosophers have criticized consequentialism because it risks exploitation or rights violation of a small number of agents for the good of the majority (one extreme example includes the justification of slavery [Kilbride, 1993]). At the same time, blind compliance with norms may also risk in developing negative consequences if the norm is defined too narrowly.

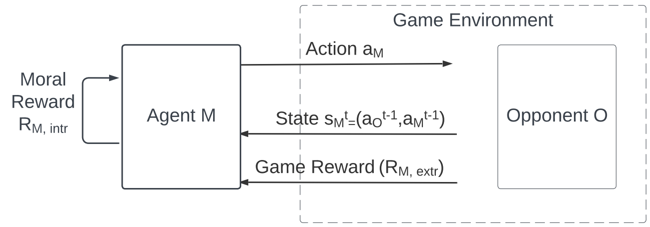

Figure 1: A step in the RL process, from the point of view of a learning moral agent M playing against a learning opponent O. Agent M calculates their moral (intrinsic) reward according to their own ethical framework, which may or may not consider the game (extrinsic) reward coming from playing the dilemma.

3 Modeling Moral Choice with Reinforcement Learning

3.1 Reinforcement Learning for Decision-Making in Social Dilemmas

Our environment is a two-player iterated dilemma game, played over 10000 iterations, which represent a single episode. At each iteration, a moral player M and an opponent O play the one-shot matrix game corresponding to one of three classic social dilemmas - Prisoner's Dilemma, Volunteer's Dilemma and Stag Hunt (see Table 1). In this Markov game [Littman, 1994], both players learn to choose actions from interacting with their opponent. At time t, the learning agent M observes a state, which is the pair of actions played by M and O at $t-1$: $s_M^t = (a_O^{t-1}, a_M^{t-1})$, and chooses an action simultaneously with the opponent O: $a_M^t, a_O^t \in \{C, D\}$. The player M then receives a reward R_M^{t+1} and observes a new state s_M^{t+1}. Both agents use tabular Q-Learning [Watkins and Dayan, 1992] to update the Q-value for a given state and action for both agents as follows:

$$Q(s^t, a^t) \leftarrow Q(s^t, a^t) + \alpha \left[R^{t+1} + \gamma \max_a Q(s^{t+1}, a) - Q(s^t, a^t) \right] \quad (1)$$

where α is the learning rate and γ is a discount factor.

We simulate moral choice using an ϵ-greedy policy, where agents act randomly ϵ% of the time and otherwise play greedily according to the Q-values learned so far. Player M can learn according to an *extrinsic* game reward $R_{M_{extr}}^{t+1}$, which depends directly on the joint actions a_M^t, a_O^t (as defined in Table 1) or according to an *intrinsic* reward $R_{M_{intr}}^{t+1}$, which we define subsequently.

3.2 Modeling Morality as Intrinsic Reward

As discussed, intrinsic rewards have been used to nudge RL agents towards learning more cooperatively in social dilem-

Agent Moral Type	*Moral Reward Function*	*Source of Morality*
Utilitarian	$R^t_{M_{intr}} = R^t_{M_{extr}} + R^t_{O_{extr}}$	External/Internal conseq.
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
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
Virtue-kindness	$R^t_{M_{intr}} = \begin{cases} \xi, & \text{if } a^t_M = C \\ 0, & \text{otherwise} \end{cases}$	Internal norm
Virtue-mixed	$R^t_{M_{intr}} = \begin{cases} \beta * (1 - \frac{\lvert R^t_{M_{extr}} - R^t_{O_{extr}} \rvert}{R^t_{M_{extr}} + R^t_{O_{extr}}}) + (1-\beta) * \hat{\xi}, & \text{if } a^t_M = C \\ \beta * (1 - \frac{\lvert R^t_{M_{extr}} - R^t_{O_{extr}} \rvert}{R_{M^t_{extr}} + R^t_{O_{extr}}}), & \text{otherwise} \end{cases}$	Internal norm & conseq.

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.

mas via certain social preferences [Hughes *et al.*, 2018; Jaques *et al.*, 2019; Peysakhovich and Lerer, 2018a]. We extend this work by defining four Q-learning moral agents M that learn according to various intrinsic rewards $R_{M_{intr}}$ (see Figure 1), and contrasting these against a traditional Q-learning *Selfish* agent[2] M who learns to maximize its extrinsic (game) reward $R_{M_{extr}}$ (e.g., [Leibo *et al.*, 2017]).

We define four different moral learners: *Utilitarian*, *Deontological*, *Virtue-equality* and *Virtue-kindness*:

- the *Utilitarian* agent tries to maximize the total payoff for both players [Bentham, 1996], and receives an intrinsic reward R^t_{intr} based on collective payoffs from this iteration, equal to $R^t_{M_{extr}} + R^t_{O_{extr}}$;
- the *Deontological* agent tries to follow the conditional cooperation norm [Fehr and Fischbacher, 2004] and gets punished by $-\xi$ for defecting against an opponent who previously cooperated (i.e., receive a negative reward $R^t_{M_{intr}} = -\xi$ if $a^{t-1}_O = C$ and $a^t_M = D$);
- the *Virtue-equality* 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* agent gets a reward ξ for cooperating against any opponent on this iteration.

Precise definitions of the reward functions are presented in Table 2. We use the same moral rewards across the three games to investigate the impact of payoff structure on learning. Given the payoff matrices of the three games used (Table 1), we set the parameter ξ in the two norm-based rewards to be $\xi = 5$, so it sends a strong signal of a value similar to the maximum payoff available in any of the three games. *Utilitarian* agents can be defined as consequentialist since their reward depends on consequences in the environment (and, in particular, on the reward of the other agent), but the source of this morality can be external (social norms) or internal (the agent's values). *Deontological* agents depend on an external norm. *Virtue-equality* agents can be considered internal consequentialist, since they follow a consequence-based internal virtue. *Virtue-kindness* agents depend on an internal norm.

Finally, we consider the fact that the true nature of virtue ethics implies a combination of virtues within a single agent (e.g., [Graham *et al.*, 2013]). Multi-objective models may be needed to represent the complexity of human values - moral or otherwise [Peschl *et al.*, 2022; Rodriguez-Soto *et al.*, 2022; Vamplew *et al.*, 2018]. Thus, we implement a *Virtue-mixed* agent as well, which uses a linear combination of the *equality* and *kindness* virtue rewards defined earlier. We use a linear combination because the two types of reward are independent - either value of the *equality* reward is possible given either value of the *kindness* reward. The formal definition of intrinsic reward for the *Virtue-mixed* agent M can be found in Table 2. We normalize the values of the *kindness* reward ξ as $\hat{\xi}$ in order to bound it to $[0, 1]$ (to match the range of the *equality* reward). The parameters β and $(1 - \beta)$, with $\beta \in [0, 1]$, define the relative weights on the two types of virtues. Here we present results for $\beta = 0.5$; an analysis of the impact of using different values for the weights can be found in Appendix F.

3.3 Measuring Social Outcomes

In order to investigate the impact of the presence of different intrinsically-motivated moral learners, we define three social outcome metrics, calculated as cumulative return values after the 10000 iterations - $G_{collective}$, G_{gini}, G_{min}:

$$G_{collective} = \sum_{t=0}^{10000} (R^t_{M_{extr}} + R^t_{O_{extr}}) \tag{2}$$

$$G_{Gini} = \sum_{t=0}^{10000} (1 - \frac{|R^t_{M_{extr}} - R^t_{O_{extr}}|}{R^t_{M_{extr}} + R^t_{O_{extr}}}) \tag{3}$$

$$G_{min} = \sum_{t=0}^{10000} \min(R^t_{M_{extr}}, R^t_{O_{extr}}). \tag{4}$$

The collective return $G_{collective}$ measures social welfare, identical to the *Utilitarian* reward accumulated over time; G_{Gini} measures the equality between rewards using the Gini

[2] We use the term moral agents 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 with the concept of rationality. For these agents $R_{M_{intr}} = R_{M_{extr}}$.

coefficient [Gini, 1912], identical to the *Virtue-equality* reward accumulated over time; and G_{min} measures the minimum reward obtained by either of the players, also accumulated over time to reflect long-term impacts.

In summary, we model pairs of agents learning against one another, and consider pairs that are characterized by the same or different moral reward functions. We explore all possible combinations of the six agents (one *Selfish* and five non-selfish intrinsically-motivated ones - *Utilitarian*, *Deontological*, *Virtue-equality*, *Virtue-kindness* and *Virtue-mixed*) in three social dilemma games (Iterated Prisoner's Dilemma, Iterated Volunteer's Dilemma, and Iterated Stag Hunt). As an additional benchmark and for potential intellectual curiosity, we also study the learning processes of these six agents against static agents from traditional Game Theory tournaments [Axelrod and Hamilton, 1981], namely *Always Cooperate*, *Always Defect*, *Tit for Tat*, and a *Random* agent. The results of these experiments are available in Appendix B.

4 Evaluation

4.1 Experimental Setup

We use a linearly decaying exploration rate ϵ (from 1.0 to 0; see Appendix E for a discussion of the effects of a smaller exploration rate), a steady learning rate $\alpha = 0.01$ (a parameter search on *Selfish* versus *Selfish* experiments demonstrated that the results were insensitive to the choice of alpha, hence we chose the smallest value), and discount factor $\gamma = 0.90$. At the start of the game, all Q-values are initialized to 0. Each pair of agents interact in one episode for 10000 iterations of a given social dilemma game. We repeat each episode 100 times, randomizing seeds and initial states at every run. All pairwise comparisons are independent.

4.2 Systematic Comparison of Dyadic Interactions

Across the three games, we investigate behaviors and outcomes emerging between each pair of agents M and O at the end of an episode using two metrics: simultaneous actions chosen (Figures 2-4) and social outcomes obtained (as defined in Equations 2-4; see Figures 5-7). An analysis of reward is presented in Appendix C.

Simultaneous Actions

For every pair of agents, we measure the proportion of action pairs (as a percentage across the 100 runs) that corresponds to mutual cooperation, one player *exploiting* the other (i.e., defecting when the other cooperates), and mutual defection. Results are presented in Figures 2-4. Visualizations of temporal dynamics are provided in Appendix A.

In the Iterated Prisoner's Dilemma (Figure 2), *Selfish* players are motivated by greed and fear. As expected [Leibo *et al.*, 2017; Sandholm and Crites, 1996], the traditional *Selfish* agent learns a 'safe' Always Defect strategy against all agents. This results in it mutually defecting against itself and *Virtue-equality* (shown in pink), and exploiting all other agents (shown in orange). For the non-selfish players, the *Utilitarian*, *Deontological*, *Virtue-kindness* and *Virtue-mixed* agents (i.e., all but *Virtue-equality*) never defect against any other agent, and as a result achieve mutual cooperation when facing one another (shown in dark green), or face 100% exploitation when facing a *Selfish* agent (shown in dark blue). The *Virtue-equality* agent, on the other hand, learns a more defensive but less cooperative strategy - it mutually defects on 100% of the runs against a *Selfish* opponent, or 50% against itself, and learns to exploit all other non-selfish agents on 15-20% of the runs (shown in light orange). Further analyses (see Appendix G) show that this is due to non-convergence of the agent's learning over the duration used in this set of experiments. Running the experiments for a higher number of steps results in slow convergence to fully cooperative strategies by the *Virtue-equality* agent, similar to other non-selfish agents, which is an interesting finding per se.

In the Iterated Volunteer's Dilemma game (Figure 3), in which players no longer fear an uncooperative partner, even the *Selfish* agent is able to avoid the least desirable outcome of mutual defection more than 75% of the time (the values in pink do not go above 25%). Furthermore, given this payoff structure, even the *Selfish* agent can achieve mutual cooperation, 21% of the time against itself, 34% against *Virtue-equality*, and over 40% against other non-selfish agents. This provides a stark contrast to the Prisoner's Dilemma, in which a *Selfish* agent could not achieve mutual cooperation at all. Two *Virtue-equality* agents playing against each other, on the other hand, mutually defect in 40% of the runs - more than against the *Selfish* agent. In contrast, the *Utilitarian*, *Deontological*, *Virtue-kindness* or *Virtue-mixed* agents never exploit an opponent (on this or the other two games - see panel in orange). However, these agents themselves might get exploited by a *Selfish* or *Virtue-equality* agent 56-57% of the time (shown in blue), though not as much as on the Prisoner's Dilemma.

Removing the greed motivation in the Iterated Stag Hunt game (Figure 4) leads to an even greater likelihood of a *Selfish* agent achieving mutual cooperation against all other agents (boosting it to 45% against *Virtue-equality*, and over 55% against other moral learners). The *Selfish* agent is also better at avoiding exploiting its opponents than in the other two games (43% maximum, shown in orange), but it is more likely to end up mutually defecting against itself (36% of the runs) or a *Virtue-equality* agent (42% of the runs, shown in pink) than in the Volunteer's Dilemma. The *Virtue-equality* agent here also converges to mutual defection against itself 48% of the time. Against other non-selfish agents, it mutually cooperates on 83% of the runs, yet still exploits them on 13% of the runs (see discussion of non-convergence in Appendix G). The *Utilitarian*, *Deontological*, *Virtue-kindness* and *Virtue-mixed* agents achieve mutual cooperation 100% of the time in the Iterated Stag Hunt, as in the other two games.

Social Outcomes

We calculate three social outcome values, as defined in Equations 2-4. Figures 5-7 present average values across 100 runs. To aid comparison between games, the color saturation varies from the possible minimum to the possible maximum cumulative value for each respective outcome in each game. Confidence Intervals around these means are presented in Appendix D - they do not change the interpretation of the results.

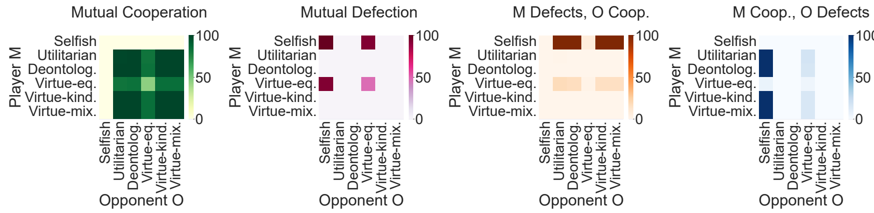

Figure 2: Iterated Prisoner's Dilemma game. Simultaneous actions played by each player M type (row) and the opponent O type (column) at the end of the learning period (10000 iterations). Action pairs are displayed as a percentage over 100 runs.

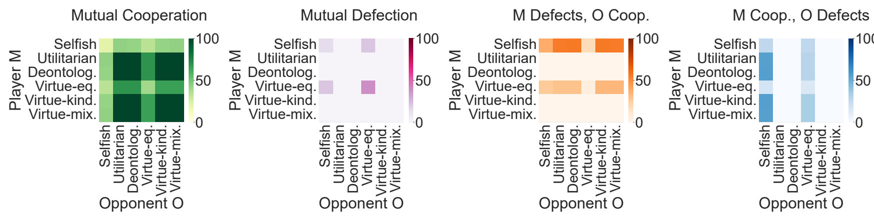

Figure 3: Iterated Volunteer's Dilemma game. Simultaneous actions played by player M type (row) and the opponent O type (column) at the end of the learning period (10000 iterations). Action pairs are displayed as a percentage over the 100 runs.

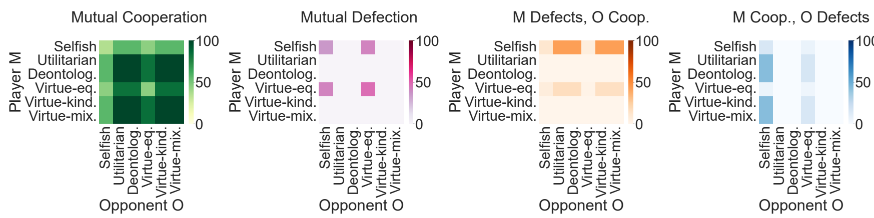

Figure 4: Iterated Stag Hunt game. Simultaneous actions played by player M type (row) and the opponent O type (column) at the end of the learning period (10000 iterations). Action pairs are displayed as a percentage over the 100 runs.

Broadly, we observe that the best social outcomes on all three games are achieved by any combination of the *Utilitarian*, *Deontological*, *Virtue-kindness* or *Virtue-mixed* agents. The *Virtue-equality* agent's strategy leads to a negative social outcome by either occasionally exploiting these four agents (this can be rectified via longer training episodes), or mutually defecting against itself or a *Selfish* opponent (as seen in Figures 2-4). The *Selfish* agent also creates negative outcomes by learning a defective policy.

We observe different outcomes depending on the characteristics of the game. In the fear-only game Stag Hunt (Figure 7), the agents achieve lower collective return (left panel) than in the other games (Figures 5-6), but greater benefit for the lowest-achieving agent (i.e., minimum return, right panel). The greatest collective return is instead obtained on the game where traditional rational agents are motivated by greed - the Volunteer's Dilemma (Figure 6, left panel). Interestingly, among non-selfish agents, the gain in collective return in this game is not achieved through exploitative actions (shown by the low values in the orange panel in Figure 3), and as a result minimum reward (right panel) stays rather high on average too. Finally, the greatest equality between agents over the 10000 iterations, as measured by the Gini return (middle panel), is also achieved on the Volunteer's Dilemma game (Figure 6), and Figure 3 shows that this mostly involves mutual cooperation (green) rather than defection (pink).

Lastly, we can consider how each agent's goals were achieved by their reward function versus others, given the parallels between the outcome metrics defined in Equations 2-4, and the moral rewards defined in Table 2. The *Utilitarian* agent, in practice, aimed to maximize collective return (left panel, Figures 5-7), and it did so better than the *Selfish* or *Virtue-equality* agent on all three games, but *Deontological*, *Virtue-kindness* and *Virtue-mixed* agents maximized collective return just as well in these environments. The *Virtue-equality* agent, which in practice aimed at maximizing Gini return (middle panel, Figures 5-7), did worse than other non-selfish agents on all three games due to non-convergence on some training runs (see Appendix G). Thus, the simple equality-based reward function was less effective at achieving equality as an outcome than other moral rewards. Finally, no agent focused directly on maximizing the least-earning agent's payoff (i.e., minimum return, right panel in Figures 5-7), but we observe that the *Selfish* and *Virtue-equality* agents achieved their goals by means of bringing another agent down, whereas all other agents were able to play morally without disadvantaging their opponent. No outcome metric directly maps to the goals of the two norm-based agents (*Deontological* and *Virtue-kindness*), but we observe that they achieve their moral goals without negative externalities to others, since they perform as well as the *Utilitarian* and *Virtue-kindness* agents on all three metrics across all games.

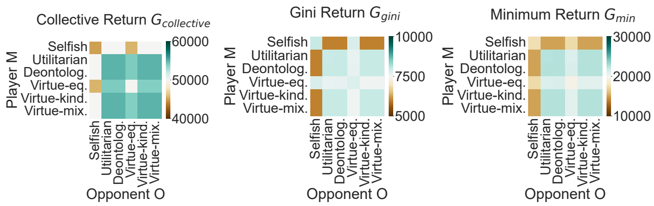

Figure 5: Iterated Prisoner's Dilemma game. Relative social outcomes observed for player type M vs. all possible opponents O. The plots display averages across the 100 runs.

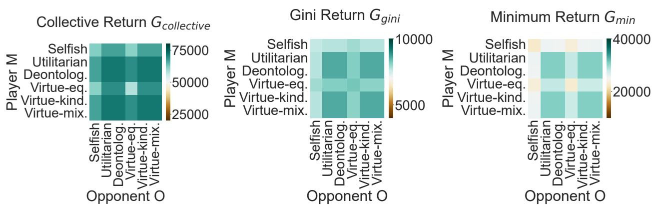

Figure 6: Iterated Volunteer's Dilemma game. Relative social outcomes observed for player type M vs. all possible opponents O. The plots display averages across the 100 runs.

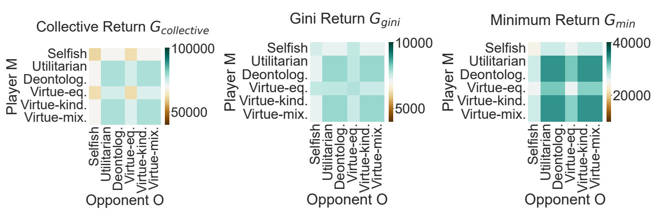

Figure 7: Iterated Stag Hunt game. Relative social outcomes observed for player type M vs. all possible opponents O. The plots display averages across the 100 runs.

Learning by a Mixed Moral Agent

On all three games, the *Virtue-mixed* agent is essentially indistinguishable from *Virtue-kindness* in terms of actions (Figures 2-4) and social outcomes obtained (Figures 5-7). Thus, an equally-weighted combination of the *equality* and *kindness* elements in a *Virtue-mixed* agent did not protect it from exploitation, but kept it just as likely as a pure-*kindness* agent to achieve the Pareto-optimal outcome of mutual cooperation. Further experiments with different weights on each virtue showed that only very high values of β (i.e., a high weight on *equality* versus *kindness*) make the mixed agent switch to more defensive and less cooperative behavior (see Appendix F for detail).

5 Discussion

First of all, we are well aware that the agents implement very simplified forms of ethical systems. Nonetheless, this work can be considered as a starting point for the implementation of more complex decision-making mechanisms with agents grounded on moral views rather than purely selfish (i.e., rational) principles. This is especially important for the design of next-generation AI systems, which might have to implement constraints derived from AI regulations or codes of practice.

In comparison with past studies of Reinforcement Learning agents acting according to certain preferences in social dilemmas, this work provides a systematic analysis of behaviors and outcomes emerging from interactions between a variety of moral agent types. This investigation has allowed us to derive insights into behavioral patterns and strategies that emerge in social dilemmas. In particular, we find that simple norm-based (*Deontological* and *Virtue-kindness*) and *Utilitarian* definitions of moral reward steer agents towards learning cooperative but easy-to-exploit policies, whereas a *Virtue-equality*-based definition is slower to converge so results in more exploitative behavior. The norm-based agents were defined in a way that promoted prosocial behavior. It remains to be seen what behavior emerges from agents who follow very different moral norms, such as punishing defecting opponents by defection.

Additionally, we find that a multi-objective reward function is able to steer a mixed virtue agent away from uncooperative behavior, but makes them more likely to get exploited. A potential development could be to study whether reward functions that mix selfish and moral objectives or implement continual curriculum learning [Bengio *et al.*, 2009] create more robust moral agents. Another interesting next step may be to model the behavior of these moral agents in more complex societies, for example with a partner selection mechanism [Anastassacos *et al.*, 2020]. Partner selection by *Utilitarian* agents may largely result in cooperative outcomes, since these agents obtain greater reward when playing against other cooperative moral agents (i.e., *Utilitarian*, *Deontological*, *Virtue-kindness* or *Virtue-mixed*). It is harder to predict the behavior of *Virtue-equality* agents with partner selection - would they drive other non-selfish agents away due to the small amount of exploitation that they perform? Finally, an extreme case would be to assume a society where norms are followed universally by all agents (an idealized situation considered, among the others, by Kant [Kant, 1981]).

6 Conclusion

In this work, we have presented for the first time a systematic investigation of the learning of RL agents whose rewards are based on moral theories using classic social dilemmas. In order to do so, we have defined reward structures that are simplified yet representative of three key ethical frameworks.

In our study, the Utilitarian, Deontological, Virtue-kindness and Virtue-mixed agents learn cooperative policies across every game. At the same time, they are exploited by the Selfish opponent. For the Virtue-equality agent, exploitative behavior emerges during the learning process before convergence. For the Virtue-mixed agent, the 'kindness' signal is stronger than 'equality'.

Our main contribution is of methodological nature: this work provides a platform on which future research can build. For example, our approach can be applied to the study of more complex societies with additional or different mechanisms, possibly characterized by a large population of agents. Given its generality, we believe it can also be used as starting point to explore interactions between human and AI agents for real-world applications.

References

[Abel *et al.*, 2016] David Abel, James MacGlashan, and Michael L. Littman. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*, 2016.

[Ajmeri *et al.*, 2020] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Elessar: Ethics in norm-aware agents. In *AAMAS'20*, pages 16–24, 2020.

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, et al. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[Anastassacos *et al.*, 2020] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *AAAI '20*, volume 34, pages 7047–7054, Apr. 2020.

[Anastassacos *et al.*, 2021] Nicolas Anastassacos, Julian García, Stephen Hailes, and Mirco Musolesi. Cooperation and reputation dynamics with reinforcement learning. In *AAMAS'21*, pages 115–123, 2021.

[Anderson *et al.*, 2005] Michael Anderson, Susan Anderson, and Chris Armen. Towards machine ethics: implementing two action-based ethical theories. In *AAAI 2005 Fall Symposium on Machine Ethics*, pages 1–7, 2005.

[Aristotle, 2009] Aristotle. *The Nicomachean Ethics*. Oxford world's classics. Oxford University Press, 2009.

[Awad *et al.*, 2018] Edmond Awad, Sohan Dsouza, Richard Kim, et al. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[Awad *et al.*, 2022] Edmond Awad, Sydney Levine, Michael Anderson, et al. Computational ethics. *Trends in Cognitive Sciences*, 26(5):388–405, 2022.

[Axelrod and Hamilton, 1981] Robert Axelrod and William D Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML'09*, pages 41–48, 2009.

[Bentahila *et al.*, 2021] Lina Bentahila, Roger Fontaine, and Valérie Pennequin. Universality and cultural diversity in moral reasoning and judgment. *Frontiers in Psychology*, 12, 2021.

[Bentham, 1996] Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation: the Collected Works of Jeremy Bentham*. Oxford University Press, 1996.

[Binmore, 2005] Ken Binmore. *Natural Justice*. Oxford University Press, 2005.

[Carroll *et al.*, 2019] Micah Carroll, Rohin Shah, Mark K Ho, et al. On the utility of learning about humans for human-ai coordination. In *NeurIPS'19*, volume 32, pages 5174–5185, 2019.

[Cavagnetto and Gahir, 2014] Stefano Cavagnetto and Bruce Gahir. Game theory - its applications to ethical decision making. *CRIS - Bulletin of the Centre for Research and Interdisciplinary Study*, 2014(1), 2014.

[Chentanez *et al.*, 2004] Nuttapong Chentanez, Andrew Barto, and Satinder Singh. Intrinsically motivated reinforcement learning. In *NeurIPS'04*, pages 1281–1288, 2004.

[Cunningham, 1967] R L Cunningham. Ethics and game theory: The prisoner's dilemma. *Papers on Non-market Decision Making*, 2(1):11–26, 1967.

[de Cote *et al.*, 2006] Enrique Munoz de Cote, Alessandro Lazaric, and Marcello Restelli. Learning to cooperate in multi-agent social dilemmas. In *AAMAS'06*, pages 783–785, 2006.

[Dignum, 2017] Virginia Dignum. Responsible autonomy. In *IJCAI'17*, pages 4698–4704, 2017.

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

[Fehr and Fischbacher, 2004] Ernst Fehr and Urs Fischbacher. Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4):185–190, 2004.

[Foerster *et al.*, 2018] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, et al. Learning with opponent-learning awareness. In *AAMAS'18*, page 122–130, 2018.

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

[Graham *et al.*, 2009] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029, 2009.

[Graham *et al.*, 2013] Jesse Graham, Jonathan Haidt, Sena Koleva, et al. Moral foundations theory: The pragmatic validity of moral pluralism. volume 47 of *Advances in Experimental Social Psychology*, pages 55–130. Academic Press, 2013.

[Harper *et al.*, 2017] Marc Harper, Vincent Knight, Martin Jones, et al. Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma. *PLOS ONE*, 12(12), 2017.

[Hegde *et al.*, 2020] Aditya Hegde, Vibhav Agarwal, and Shrisha Rao. Ethics, prosperity, and society: moral evaluation using virtue ethics and utilitarianism. In *IJCAI'20*, pages 167–174, 2020.

[Hofbauer and Sigmund, 1998] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[Hughes *et al.*, 2018] Edward Hughes, Joel Z Leibo, Matthew Phillips, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. In *NeurIPS'18*, pages 3330–3340, 2018.

[Jaques *et al.*, 2019] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, et al. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *ICML'19*, pages 3040–3049, 2019.

[Kant, 1981] Immanuel Kant. Grounding for the metaphysics of morals. 1785. *Trans. James W. Ellington. Indianapolis: Hackett*, 1981.

[Kilbride, 1993] Daniel Kilbride. Slavery and utilitarianism: Thomas cooper and the mind of the old south. *The Journal of Southern History*, 59(3):469–486, 1993.

[Kleiman-Weiner *et al.*, 2017] Max Kleiman-Weiner, Rebecca Saxe, and Joshua B. Tenenbaum. Learning a commonsense moral theory. *Cognition*, 167:107–123, 2017. Moral Learning.

[Leibo *et al.*, 2017] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, et al. Multi-agent reinforcement learning in sequential social dilemmas. In *AAMAS'17*, page 464–473, 2017.

[Littman, 1994] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *ICML'94*, page 157–163, 1994.

[Mabille and Stoker, 2021] Louise Mabille and Henk Stoker. *The Morality Wars: The Ongoing Debate Over the Origin of Human Goodness*. Fortress Academic, 2021.

[McKee *et al.*, 2020] Kevin R. McKee, Ian Gemp, Brian McWilliams, et al. Social diversity and social preferences in mixed-motive reinforcement learning. In *AAMAS'20*, page 869–877, 2020.

[Peschl *et al.*, 2022] Markus Peschl, Arkady Zgonnikov, Frans A. Oliehoek, and Luciano C. Siebert. MORAL: aligning AI with human norms through multi-objective reinforced active learning. In *AAMAS'22*, page 1038–1046, 2022.

[Peysakhovich and Lerer, 2018a] Alexander Peysakhovich and Adam Lerer. Consequentialist conditional cooperation in social dilemmas with imperfect information. In *ICLR'18*, 2018.

[Peysakhovich and Lerer, 2018b] Alexander Peysakhovich and Adam Lerer. Prosocial learning agents solve generalized stag hunts better than selfish ones. In *AAMAS'18*, page 2043–2044, 2018.

[Poundstone, 1993] William Poundstone. *Prisoner's Dilemma: John von Neumann, Game Theory, and the Puzzle of the Bomb*. Anchor Books, 1993.

[Rahwan *et al.*, 2019] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.

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

[Rodriguez-Soto *et al.*, 2021] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodriguez Aguilar. Multi-objective reinforcement learning for designing ethical environments. In *IJCAI'21*, pages 545–551, 2021.

[Rodriguez-Soto *et al.*, 2022] Manel Rodriguez-Soto, Marc Serramia, Maite Lopez-Sanchez, and Juan Antonio Rodriguez-Aguilar. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24(1):1–17, 2022.

[Salahshour, 2022] Mohammad Salahshour. Interaction between games give rise to the evolution of moral norms of cooperation. *PLOS Computational Biology*, 18(9):1–35, 2022.

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

[Sigmund, 2010] Karl Sigmund. *The Calculus of Selfishness*. Princeton University Press, 2010.

[Skyrms, 2001] Brian Skyrms. The stag hunt. *Proceedings and Addresses of the American Philosophical Association*, 75(2):31–41, 2001.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement Learning, Second Edition: An Introduction*. MIT Press, 2018.

[Tolmeijer *et al.*, 2021] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, et al. Implementations in machine ethics: a survey. *ACM Computing Surveys*, 53(6), 2021.

[Vamplew *et al.*, 2018] Peter Vamplew, Richard Dazeley, Cameron Foale, et al. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20(1):27–40, 2018.

[Wallach and Allen, 2009] Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, feb 2009.

[Wallach, 2010] Wendell Wallach. Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics and Information Technology*, 12(3):243–250, 2010.

[Watkins and Dayan, 1992] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

[Yu *et al.*, 2018] Han Yu, Zhiqi Shen, Chunyan Miao, et al. Building ethics into artificial intelligence. In *IJCAI'18*, pages 5527–5533, 2018.