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

A Simultaneous Pairs of Actions over Time - Learning Player vs Learning Opponent

In the paper we present simultaneous actions played on the final iteration, after the learning period of 10000 iterations is complete, and when is there is no longer any exploration ($\epsilon = 0$). In Figures 1-3 we present the simultaneous actions played over time, to show the dynamics of learning for every pair of agents.

Figure 1: Iterated Prisoner's Dilemma game. Simultaneous pairs of actions observed over time. Learning player M (row) vs. learning opponent O (column).

Figure 2: Iterated Volunteer's Dilemma game. Simultaneous pairs of actions observed over time. Learning player *M* (row) vs. learning opponent *O* (column).

B Learning Against Static (Baseline) Agents - Results

B.1 Simultaneous Pairs of Actions over Time - Learning Player vs Static Opponent

In Figures 4-6 we present simultaneous actions over time for learning agents versus static opponents - *Always Cooperate, Always Defect, Tit for Tat and Random*. These were run as a benchmark before implementing learning pairs of agents - but provide clear insights into the behavior of our moral agents against predictable opponents whose behavior is stable.

B.2 Summary of Simultaneous Actions - Learning Player vs Static Opponent

In this section we provide a summary of the simultaneous actions performed on the last iteration against static (predictable) opponents.

Considering learning against static opponents on the Iterated Prisoner's Dilemma (Figure 7), the traditional *Selfish* agent learns to Defect on 100% of the runs against everyone. The *Virtue-equality* agent learns efficiently against *Always Cooperate* and defends itself against exploitation by *Always Defect*. The *Utilitarian*, *Virtue-kindness* and *Virtue-mixed* agents also learn efficiently against *Always Cooperate*, but do not protect themselves from exploitation by an *Always Defect* agent (see exploitation in blue). The *Deontological* agent is able to defend itself somewhat better against an *Always Defect* agent (by achieving mutual defection on half of the runs) because its reward function allows it to play randomly against a defector. Against the reciprocal *Tit for Tat*, most moral agents learn to always cooperate, but the *Virtue-equality* agent converges to 50% mutual defection. This once again highlights that a dyadic interaction between agents focused on equality (*Virtue-equality*), reciprocity (*Tit for Tat*) or maximizing their own game reward (*Selfish*) can end up in the inefficient equilibrium. Finally, against a *Random* agent most

Figure 3: Iterated Stag Hunt game. Simultaneous pairs of actions observed over time. Learning player M (row) vs. learning opponent O (column).

moral agents implement a 'safe' *Always Cooperate* strategy (which results in them being exploited half of the time) - except the equality agent, which plays a random strategy against the Random opponent and thus gets exploited less frequently.

Learning with moral rewards against static agents in the Iterated Volunteer's Dilemma (Figure 8) results in pairs of actions similar to the Iterated Prisoner's Dilemma. The equality agent is now the only one to end up in the inefficient mutual defection situation - 100% of the time against *Always Defect*, and half the time and a quarter of the time against *Tit for Tat* or *Random* respectively. The *Selfish* agent here is more likely to end up in an exploitative situation - either it exploits an *Always Cooperate* agent (see orange), or it gets exploited by *Always Defect* (see blue), or half-half against *Tit for Tat*. Against a *Random* agent the *Selfish* learner now learns a *Random* strategy, instead of an *Always Defect* strategy as in the Iterated Prisoner's Dilemma.

Finally, learning against static opponents in the Iterated Stag Hunt game (Figure 9), all agents including the *Selfish* one converge to the Pareto-optimal mutual cooperation against *Always Cooperate*. Against *Always Defect*, once again the *Utilitarian*, *Virtue-kindness* and *Virtue-mixed* agents are defenseless (and get exploited 100% of the time - see plots in blue), the *Deontological* agent is protected from exploitation half of the time because they play randomly against a defector, and the *equality* and *Selfish* agents are able to achieve 100% mutual defection. *Tit for Tat* elicits mutual defection half of the time from *Virtue-equality* and *Selfish* agents - and the other half the time it elicits mutual cooperation.

B.3 Summary of Reward - Learning Player vs Static Opponent

Figure 10 visualizes the game and moral reward obtained by each moral player against all static agent types - *Always Cooperate*, *Always Defect*, *Tit for Tat* and *Random*.

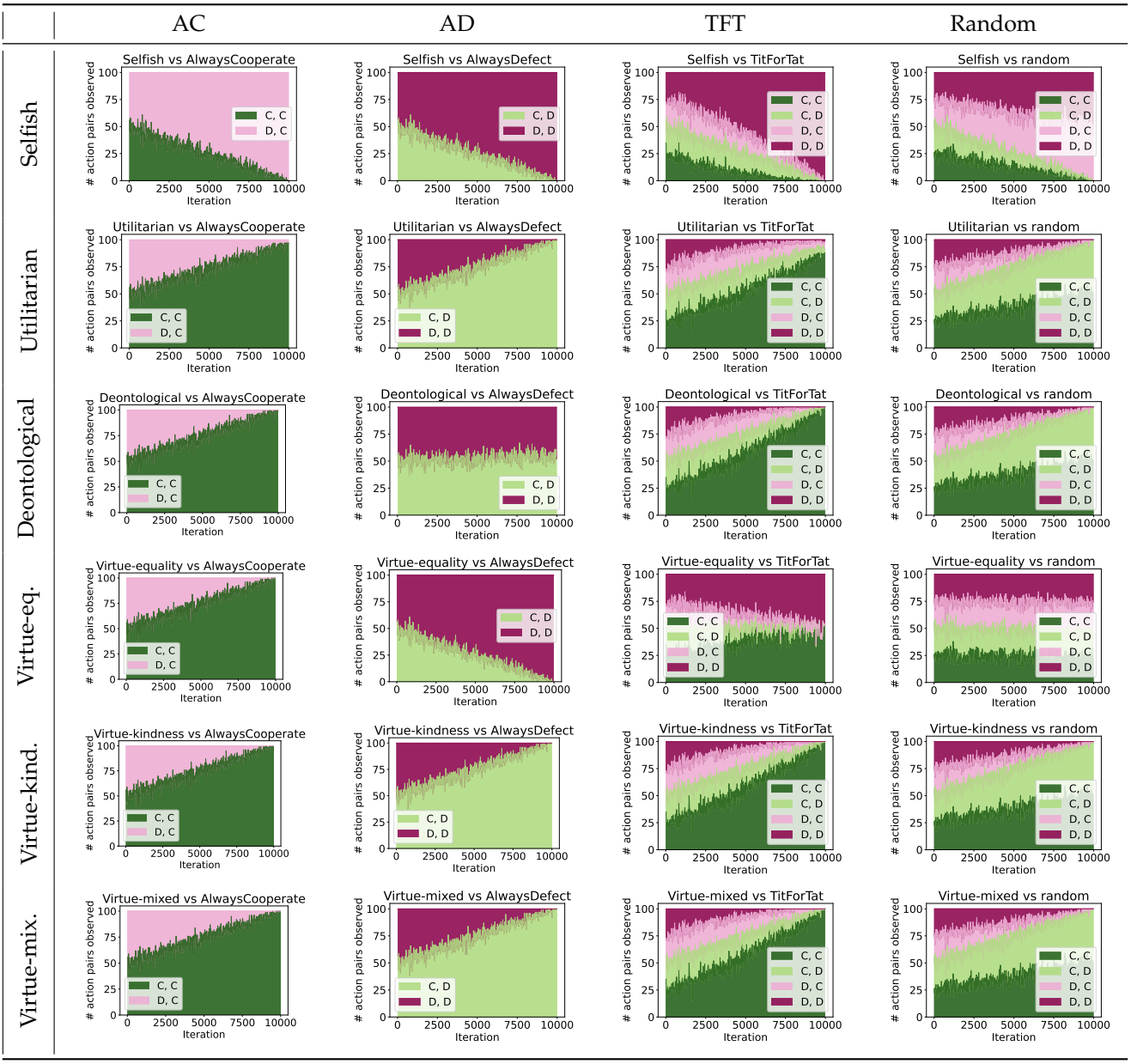


Figure 4: Iterated Prisoner's Dilemma game. Simultaneous pairs of actions observed over time. Learning player M (row) vs. static opponent O (column).

B.4 Summary of Social Outcomes - Learning Player vs Static Opponent

In Figures 11-13 we provide a summary of the social outcomes obtained when moral players learn against static opponents.

C Game and Moral Reward

C.1 Reward obtained by the end of the training

Next, we consider *game reward* (i.e., in our case, equal to the extrinsic reward) and *moral reward* (i.e., in our case, equal to the intrinsic reward) accumulated by each agent type M against all other agent types. Figure 14 presents average cumulative rewards across 100 runs, and their Confidence Intervals, for all three games.

We first analyze extrinsic game reward obtained (row one in panels A,B,C, Figure 14). Across all three envi-

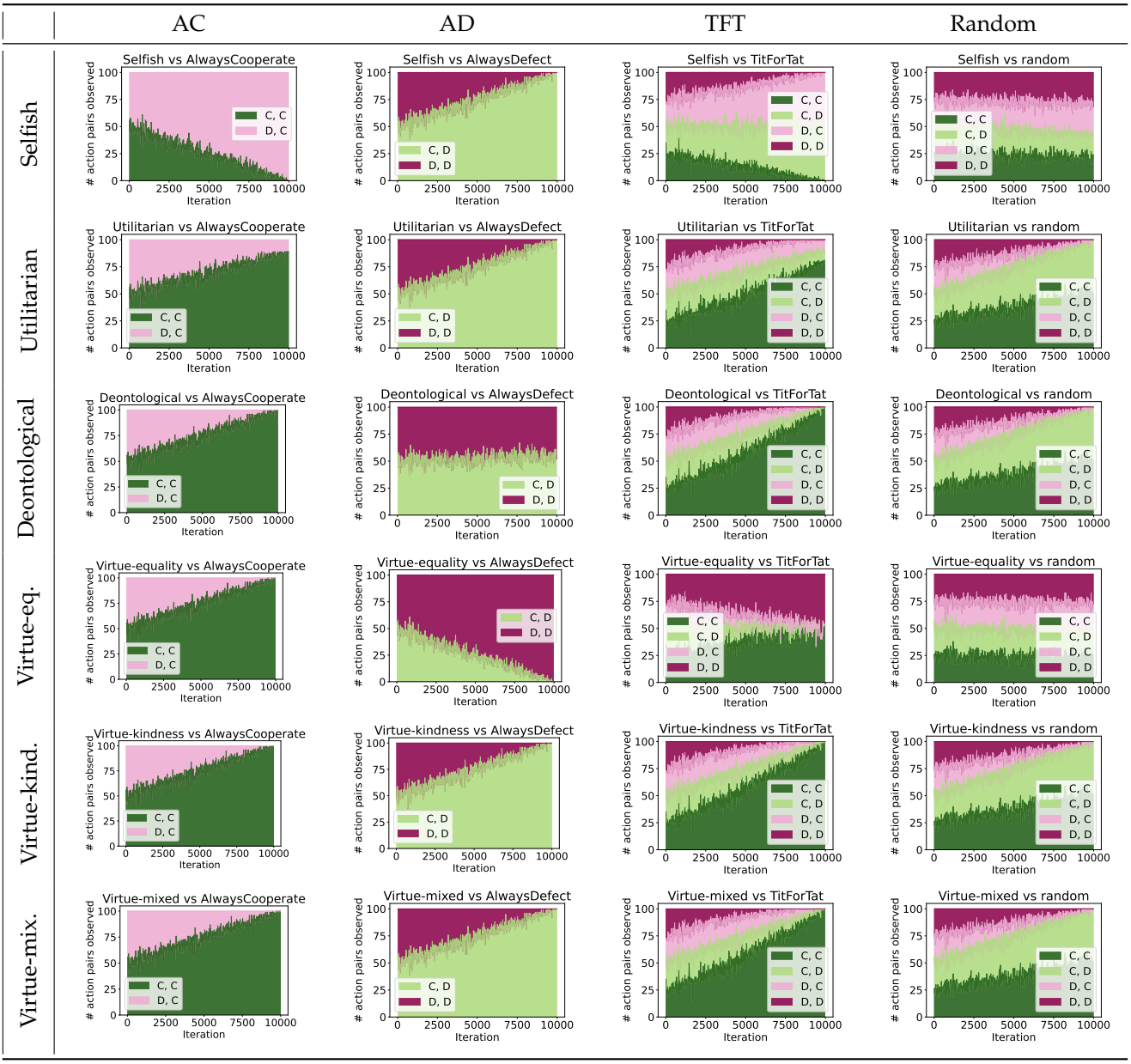


Figure 5: Iterated Volunteer's Dilemma game. Simultaneous pairs of actions observed over time. Learning player M (row) vs. static opponent O (column).

ronments, we observe that, on average, the *Selfish* agent (first column in each panel) shows better performance in games against most non-selfish opponents (due to the exploitation observed), but significantly worse when facing an agent of their same type or the *Virtue-equality* agent (due to the mutual defection observed. We also see that the *Utilitarian*, *Deontological*, *Virtue-kindness* and *Virtue-mixed* agents similarly obtain the highest game reward when facing another non-selfish agent of this type - because of mutual cooperation (as observed in the pairwise actions). However, against *Selfish* or *Virtue-equality* agents, they do worse on the three games, since exploitation emerges.

Considering the intrinsic moral reward (row two in panels A,B,C, Figure 14), we observe that across the three environments the moral agents are broadly able to learn to achieve high intrinsic reward as expected. However, it is worth noting some differences that can be observed in the figure. Specifically, the *Utilitarian*, *Deontological* and *Virtue-mixed* agents obtain a smaller moral reward when learning against *Selfish* or *Virtue-equality* opponents. This is a direct consequence of the alignment between the best playing strategies for the game and those that emerge from acting morally. The *Virtue-equality* or *Virtue-kindness* agents achieve stable levels of reward regardless of who they learn against - so exploiting or being exploited by others, as observed in the pairwise action plots, does not

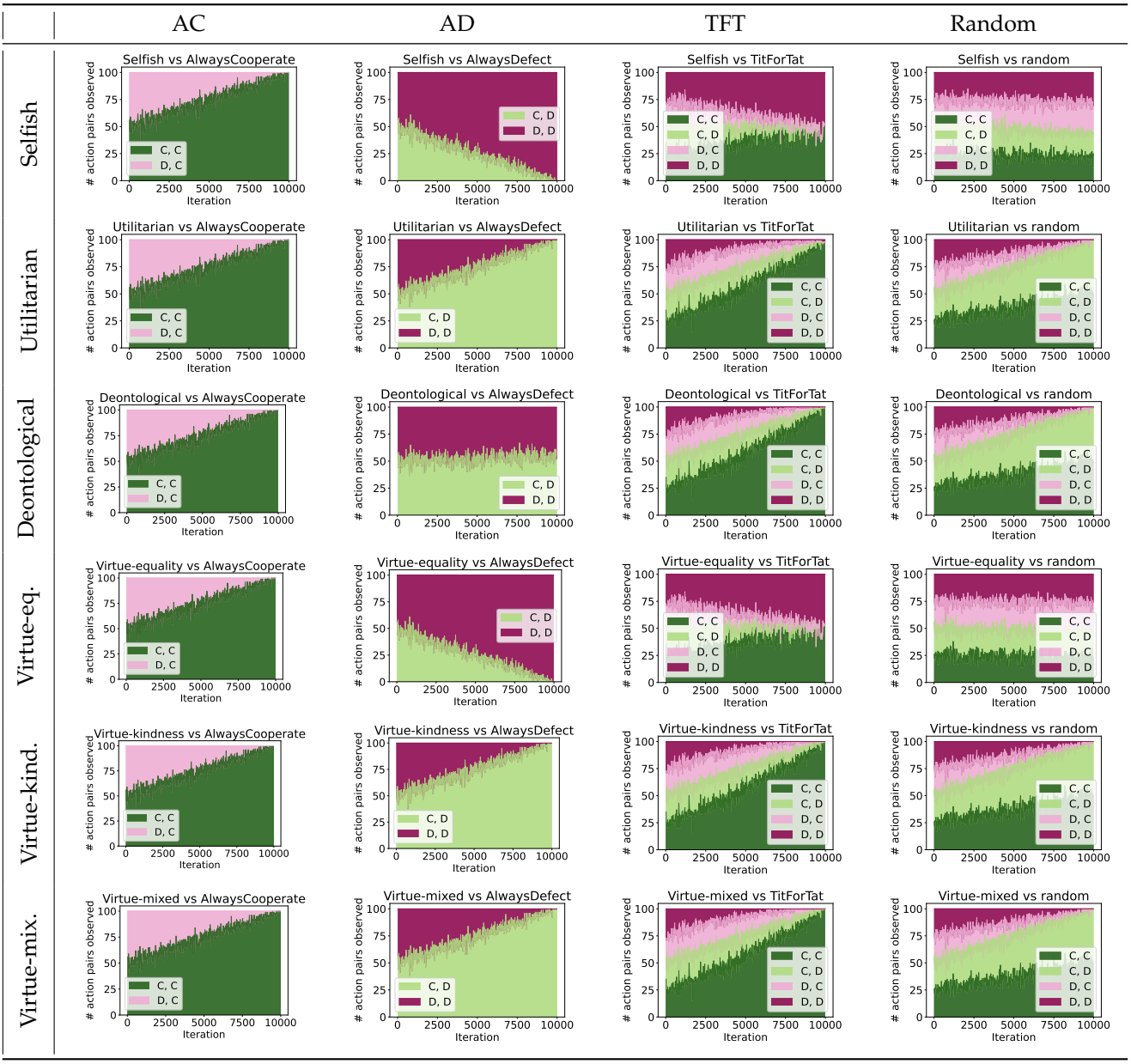


Figure 6: Iterated Stag Hunt game. Simultaneous pairs of actions observed over time. Learning player M (row) vs. static opponent O (column).

get reflected in the average cumulative reward for these agents.

Furthermore, for the *Utilitarian*, *Deontological* and *Virtue-mixed* players (second, third and final column in panels A,B,C, Figure 14), we observe an interaction between moral and game reward. We find that playing better morally is associated with better game performance - for example, these agents obtain a smaller moral reward against *Selfish* or *Virtue-equality* opponents, and they also do worse in terms of game reward on these same occasions. No such effects are observed for the *Virtue-equality* and *Virtue-kindness* agents on any of the games.

C.2 Reward over time

In the main paper we show cumulative game and moral reward obtained by learning player M vs. all possible learning opponents O. Here we present the same rewards over time (over the 10000 iterations), for a consideration of the learning dynamics (Figures 15-17). Due to the linear ϵ-decay from 1.0 to 0, we observe a near-linear convergence to the final outcome over time for all types of agents, with only slight variations in the shape. We also

observe an overlap in the learning curves of the *Utilitarian, Deontological, Virtue-kindness* and *Virtue-mixed* agents.

D Social Outcomes - with Confidence Intervals

In the paper we present heat-maps summarizing average values (across 100 runs) for collective, Gini and minimum return on all three games. In Figures 18 -20 we present the same data but in bar plots showing 95% Confidence Intervals. As stated in the paper, a consideration of Confidence Intervals does not change the interpretation of the relative return values.

E Does Exploration Aid Moral Agents' Learning?

In the main body of the paper we present agents who start off by exploring 100% of the time and then linearly decay their exploration rate to 0 by the final iteration. This allows the agents to observe all state-action pairs enough times early in the learning process, and they learn to play optimally (i.e. maximizing their moral reward $R_{M_{intr}}$ by the end of the 10000 iterations). It may be of interest to understand how our moral agents might learn without such major exploration at the start.

In Figure 21 we present the impact of implementing a less exploratory agent - one that starts off exploring 5% of the time and maintains a steady exploration rate instead. For ease of interpretation, we consider the case of each agent learning against its own kind on the Iterated Prisoner's Dilemma (patterns of learning are similar across the three games).

With smaller exploration (left, $\epsilon = 5\%$), the *Selfish*, *Utilitarian* or *Virtue-equality* agents do not learn a consistent strategy across the 100 runs. The *Selfish* agent learns three strategies - each being learned on $1/3$ of the runs: mutual defection, to exploit its opponent, or to be exploited. This is sub-optimal learning, as the *Selfish* agent that is maximizing its own game reward would have gotten a greater payoff from never being exploited. The two consequentialist agents - i.e. *Utilitarian* and *Virtue-equality* - learn one of the four possible strategies strategies on $1/4$ of the runs each. We observe that this learning stabilizes early on, and a deeper analysis showed that it is heavily impacted by early experience - i.e. the agents update their Q-values quickly at the start and then remain stuck at a local optimum and unable to learn the optimal strategy over the 10000 iterations.

The high exploration rate (right, $\epsilon = 100\%$ decaying to 0), on the other hand, allows the agents to observe the value of the alternative action and/or state, and to learn more optimally. The *Deontological* or *Virtue-kindness* agents learned to mutually cooperate in either of the settings, so exploration rate has less impact on their learning.

F The Effect of Different Weights on the two Virtues in *Virtue-mixed* Agent.

In the main results, we found that the *Virtue-mixed* agent with equivalent *equality* and *kindness* weights ($\beta = 0.5$) learned to be exploited as much as the *Virtue-kindness* agent (equivalent to $\beta = 0$), as demonstrated by *Virtue-mixed* learning against all moral opponents (presented in the paper and in pairwise action plots above - Figures 1-6). To investigate this further, in Figures 22-24 we explore what proportions of *equality* versus *kindness* in the multi-objective reward allow the *Virtue-mixed* agent to defend themselves better against exploitation. Across all three games we find that only very large relative weightings on the *equality* reward ($\beta > 0.8$) steer the mixed agent away from being exploited, but as a result the mixed agent essentially behaves as a purely *equality*-driven agent would.

G Why Do Equality-focused Agents Learn to Exploit their Opponent?

We note in the main paper that, on the final iteration, the *Virtue-equality* agent learn to exploit other non-selfish agents a small proportion of the time (up to 20%). To investigate this further and understand the underlying types of strategies that the equality agent learned, we investigate an example case of *Virtue-equality* learning against a *Utilitarian* agent on the Iterated Prisoner's Dilemma. (We choose the *Utilitarian* agent as an example of one of the most cooperative ones). We visualize the last 20 actions performed by every player in the pair (see Figure 25). We can observe that over the 100 runs, the equality agent learn to play a mixed strategy that alternates between D|(C,C) and C|(C,C) 11% of the time, and also learn an exploitative Always Defect strategy (D|(C,C)) on 9% of the runs. These runs are not efficient in terms of reward obtained given that the opponent iw an always-cooperative *Utilitarian* agent, so the best response for *Virtue-equality* would have been to cooperate against them to get the best equality score.

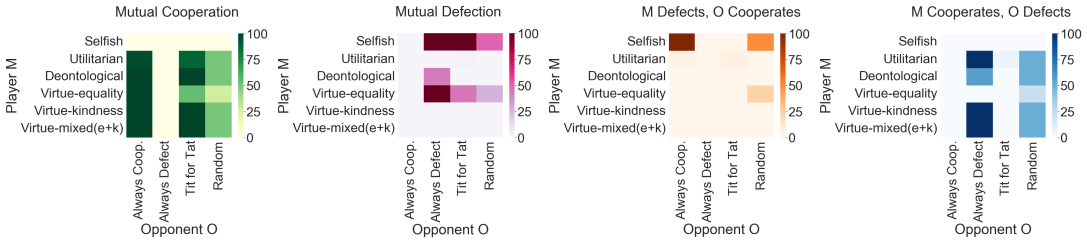


Figure 7: Iterated Prisoner's Dilemma game. Simultaneous actions played by player M type and the static opponent O type at the end of the learning period (10000 iterations). Action pairs are displayed as a percentage over the 100 runs.

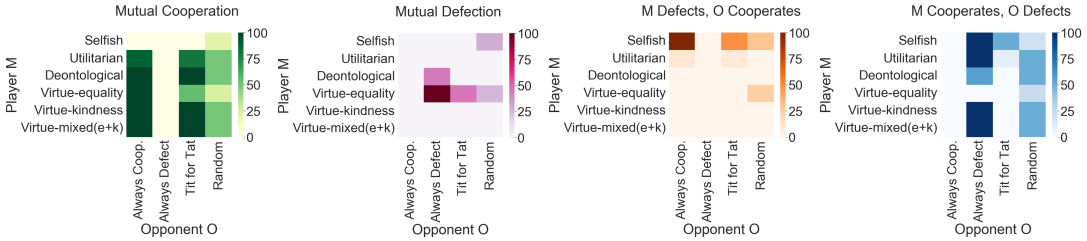


Figure 8: Iterated Volunteer's Dilemma game. Simultaneous actions played by player M type and the static opponent O type at the end of the learning period (10000 iterations). Action pairs are displayed as a percentage over the 100 runs.

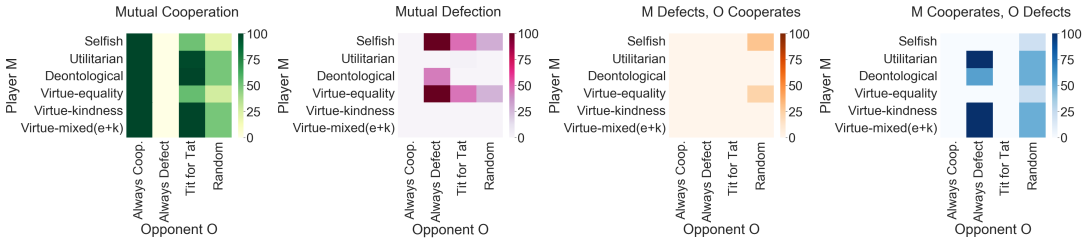
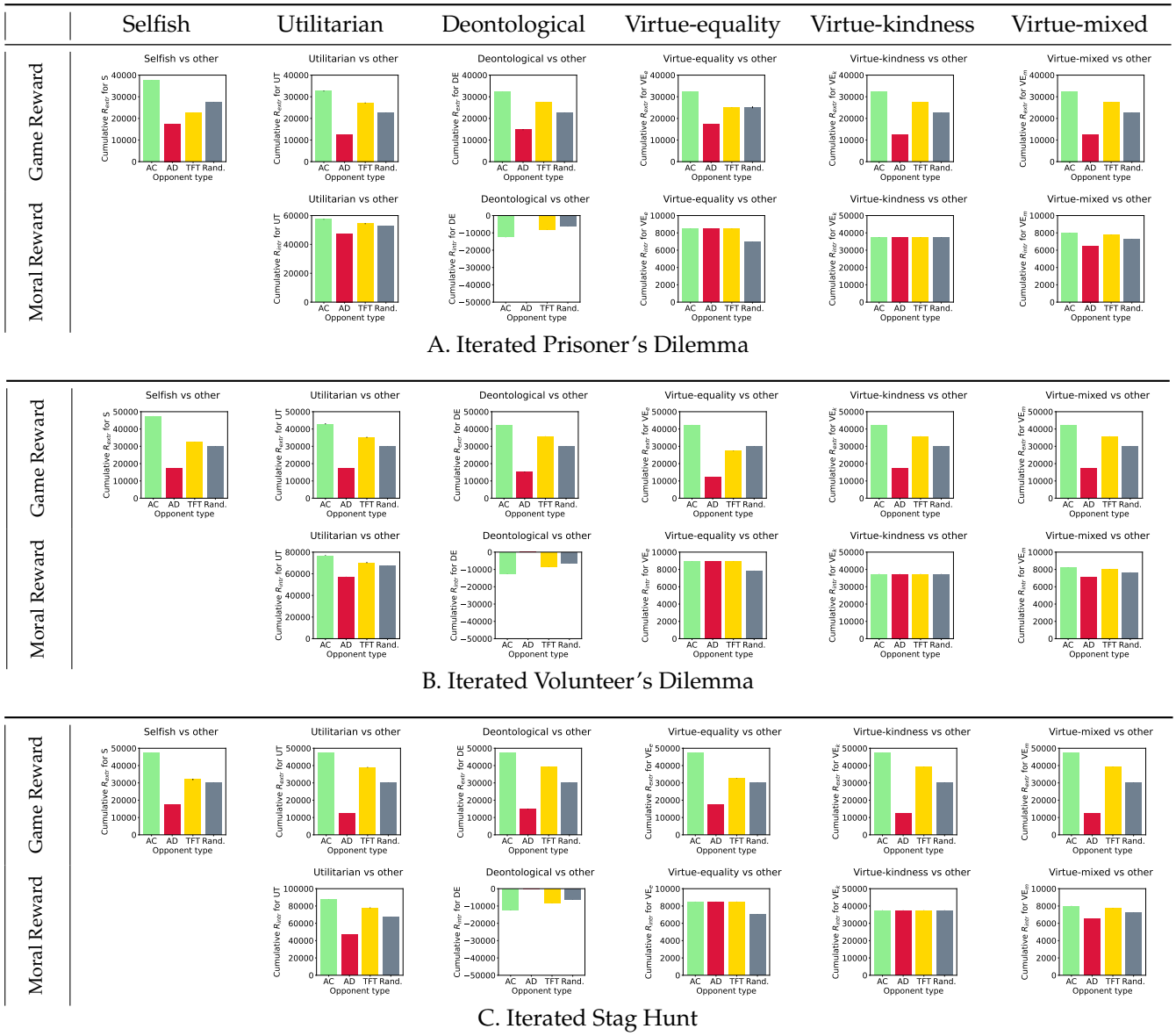


Figure 9: Iterated Stag Hunt game. Simultaneous actions played by player M type and the static opponent O type at the end of the learning period (10000 iterations). Action pairs are displayed as a percentage over the 100 runs.



A. Iterated Prisoner's Dilemma

B. Iterated Volunteer's Dilemma

C. Iterated Stag Hunt

Figure 10: Game & moral reward (cumulative) obtained after 10000 iterations by a given player type M (column) vs. all possible static opponents O - for all three games (panels A-C). The plots display averages across the 100 runs $\pm$ 95%CI.

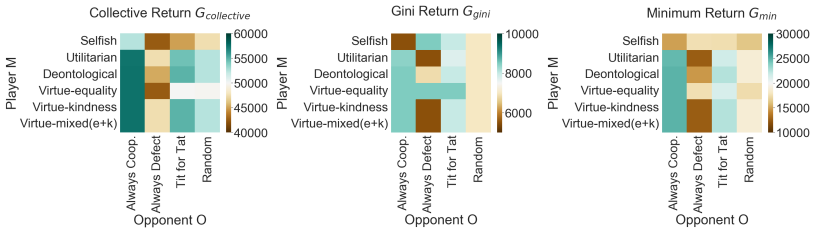


Figure 11: Iterated Prisoner's Dilemma game. Relative social outcomes observed after 10000 iterations for learning player type M (row) vs. all possible static opponents O. The plots display averages across the 100 runs.

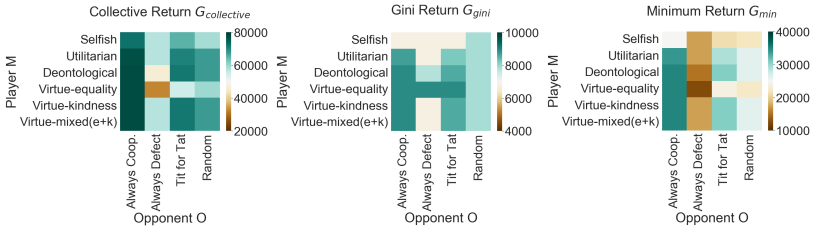


Figure 12: Iterated Volunteer's Dilemma game. Relative social outcomes observed after 10000 iterations for learning player type M (row) vs. all possible static opponents O. The plots display averages across the 100 runs.

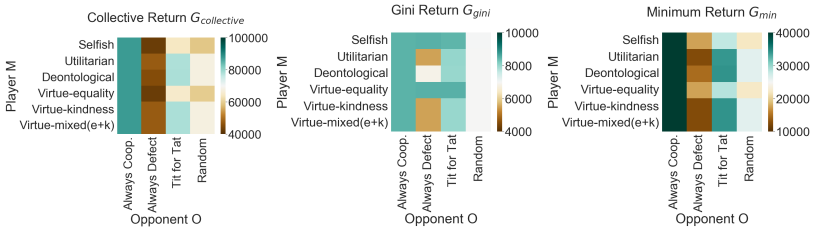


Figure 13: Iterated Stag Hunt game. Relative social outcomes observed after 10000 iterations for learning player type M (row) vs. all possible learning static O. The plots display averages across the 100 runs.

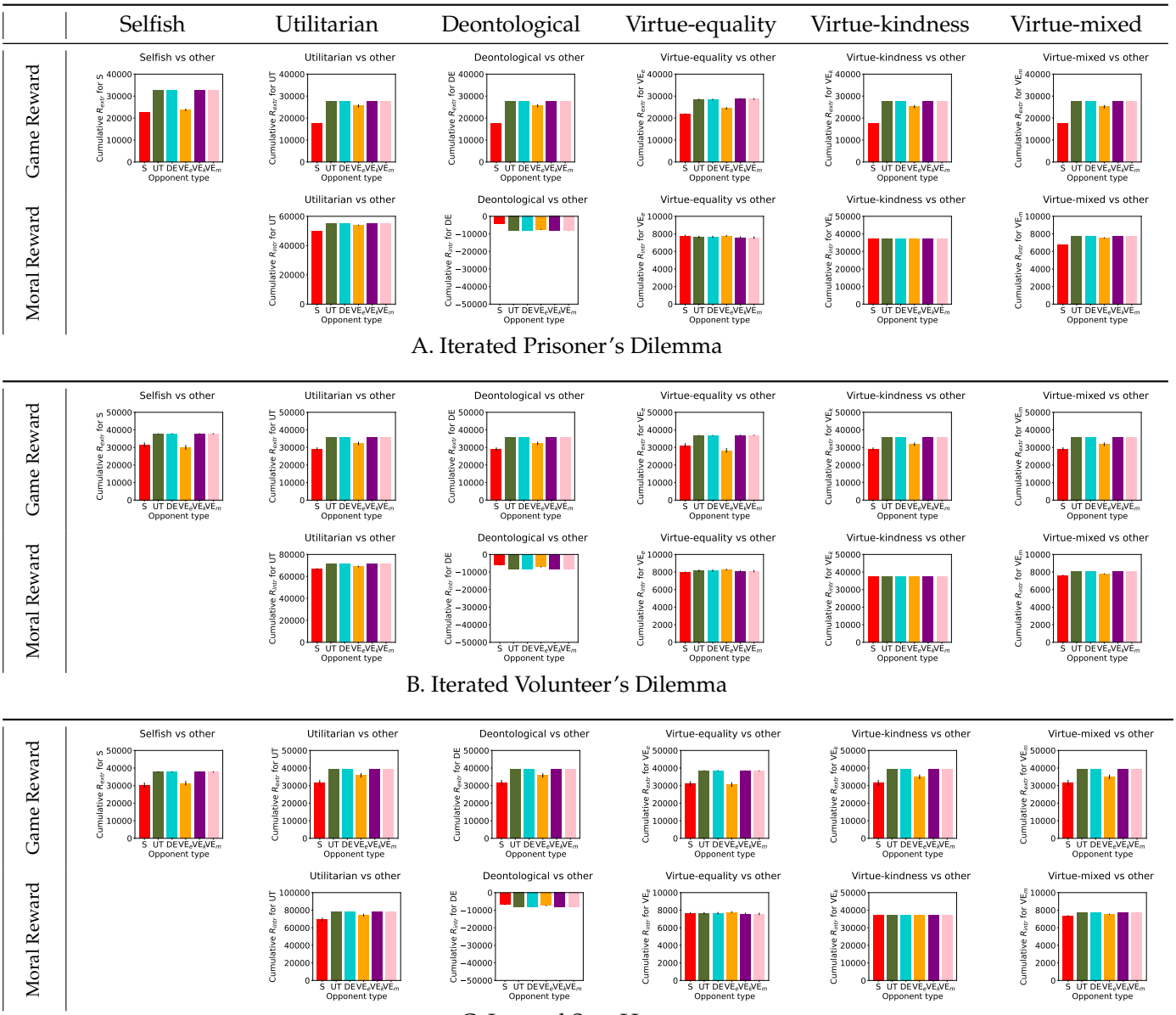


Figure 14: Game & moral reward (cumulative) obtained after 10000 iterations by a given player type M (column) vs. all possible learning opponents O - for all three games (panels A-C). The plots display averages across the 100 runs $\pm$ 95%CI.

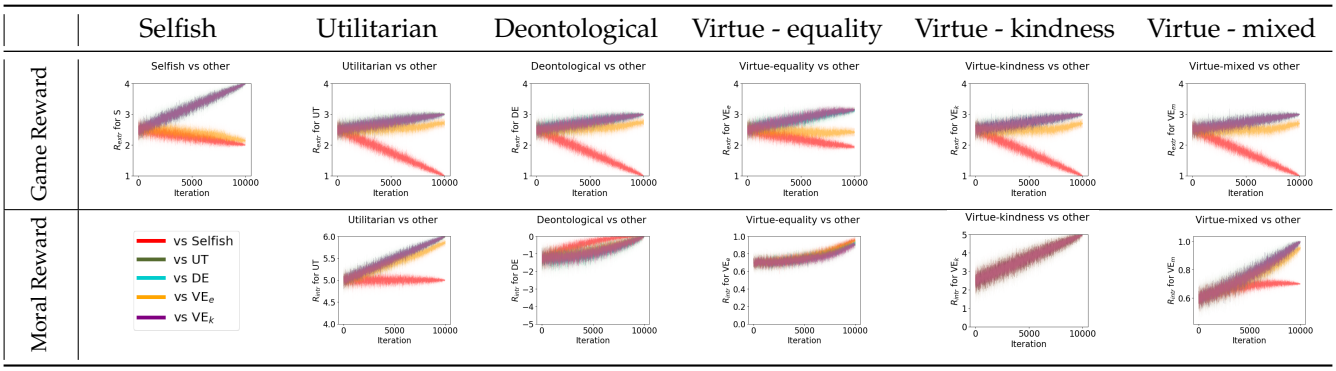


Figure 15: Iterated Prisoner's Dilemma game. Game & moral reward (per iteration) obtained by moral learning player type M (row) vs. all possible learning opponents O. The plots display average across the 100 runs $\pm$ 95%CI.

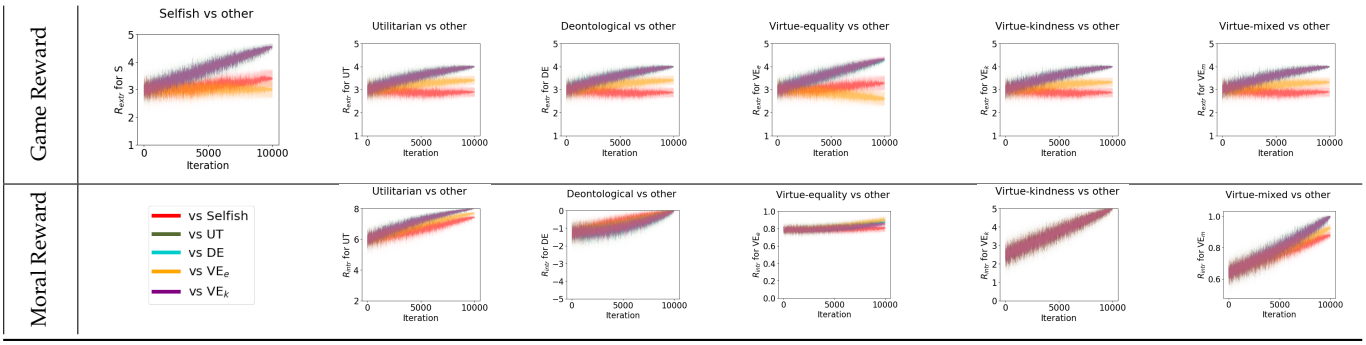


Figure 16: Iterated Volunteer's Dilemma game. Game & moral reward (per iteration) obtained by moral learning player type M (row) vs. all possible learning opponents O. The plots display average across the 100 runs $\pm$ 95%CI.

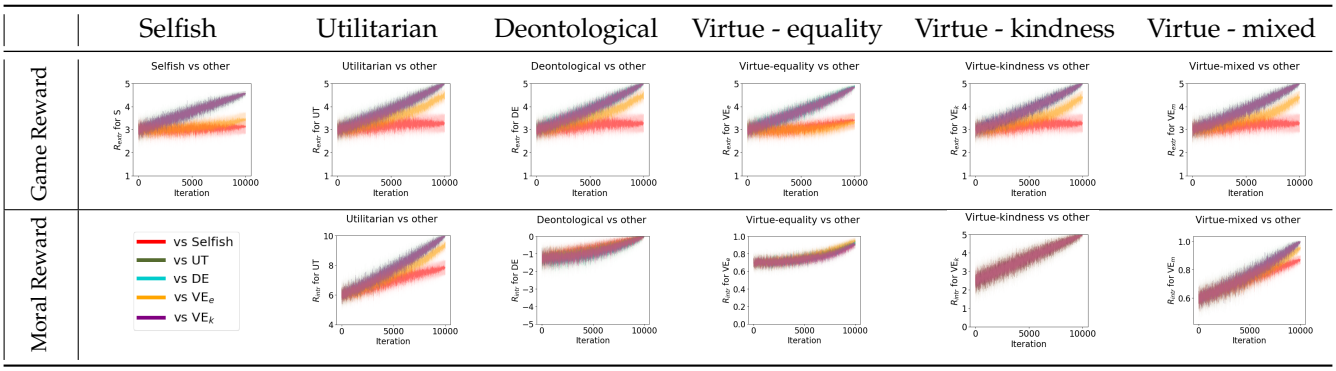


Figure 17: Iterated Stag Hunt game. Game & moral reward (per iteration) obtained by moral learning player type M (row) vs. all possible learning opponents O. The plots display average across the 100 runs $\pm$ 95%CI.

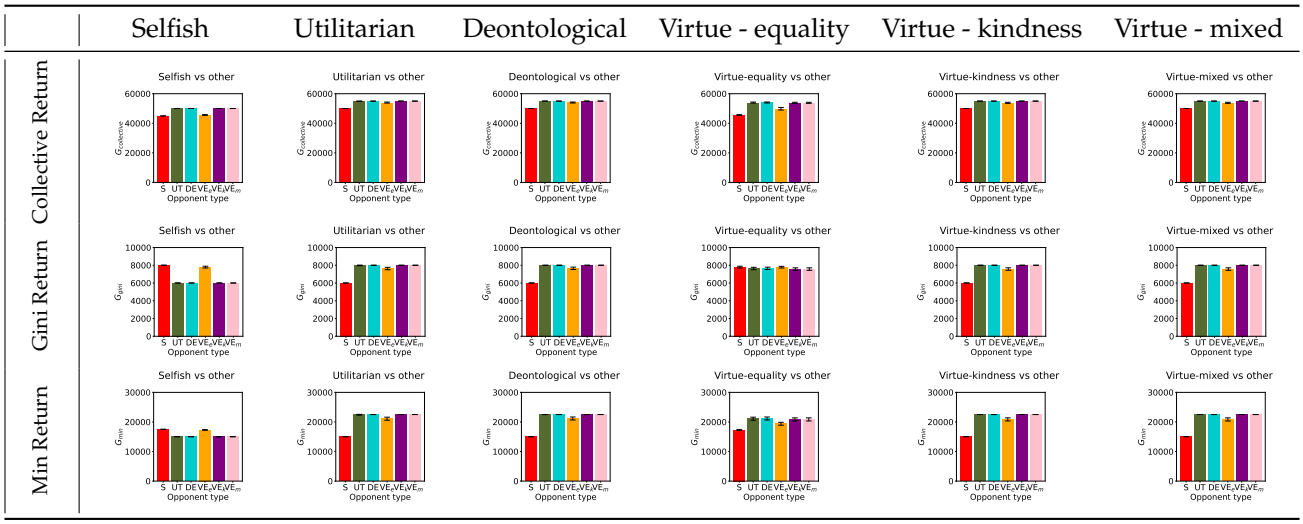


Figure 18: Iterated Prisoner's dilemma game. Relative societal outcomes observed for learning player type M (row) vs. all possible learning opponents O. The plots display averages across the 100 runs $\pm$ 95%CI.

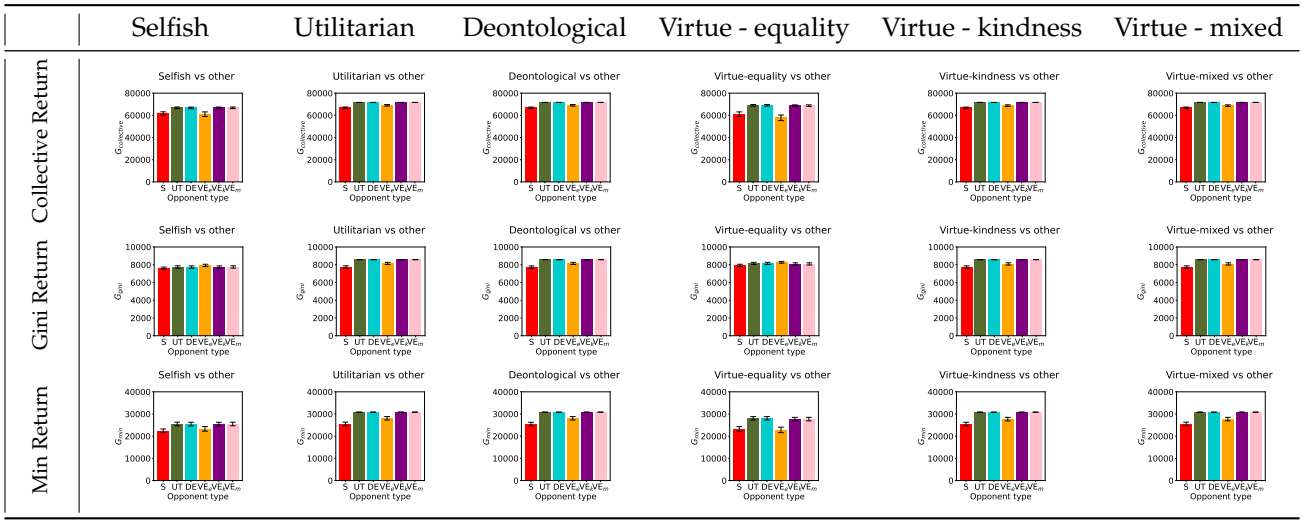


Figure 19: Iterated Volunteer's dilemma game. Relative societal outcomes observed for learning player type M (row) vs. all possible learning opponents O. The plots display averages across the 100 runs $\pm$ 95%CI.

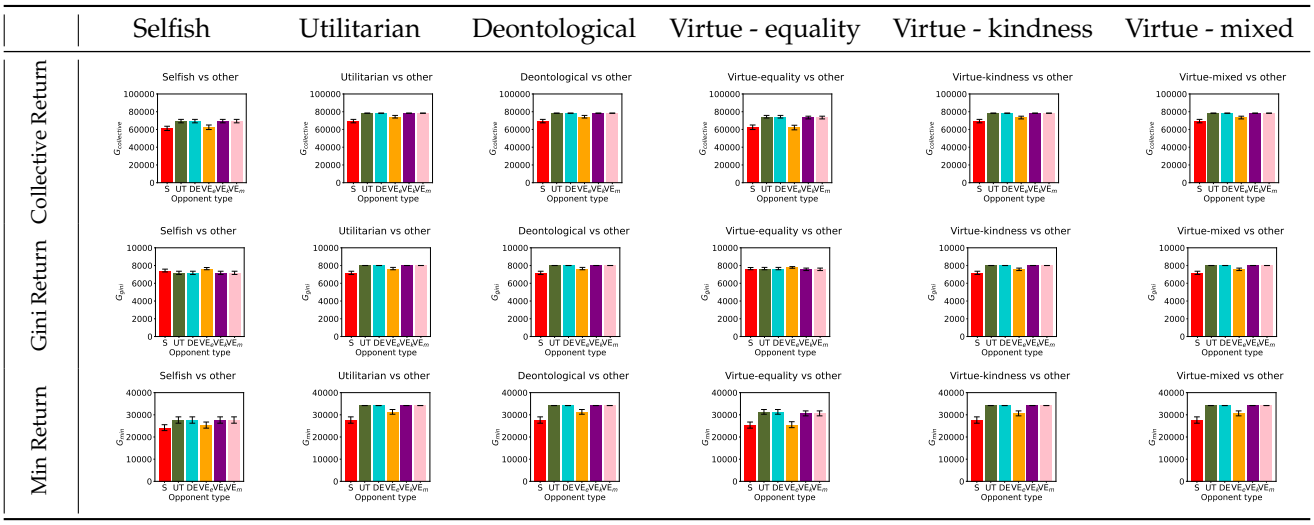


Figure 20: Iterated Stag Hunt game. Relative societal outcomes observed for learning player type M (row) vs. all possible learning opponents O. The plots display averages across the 100 runs $\pm$ 95%CI.

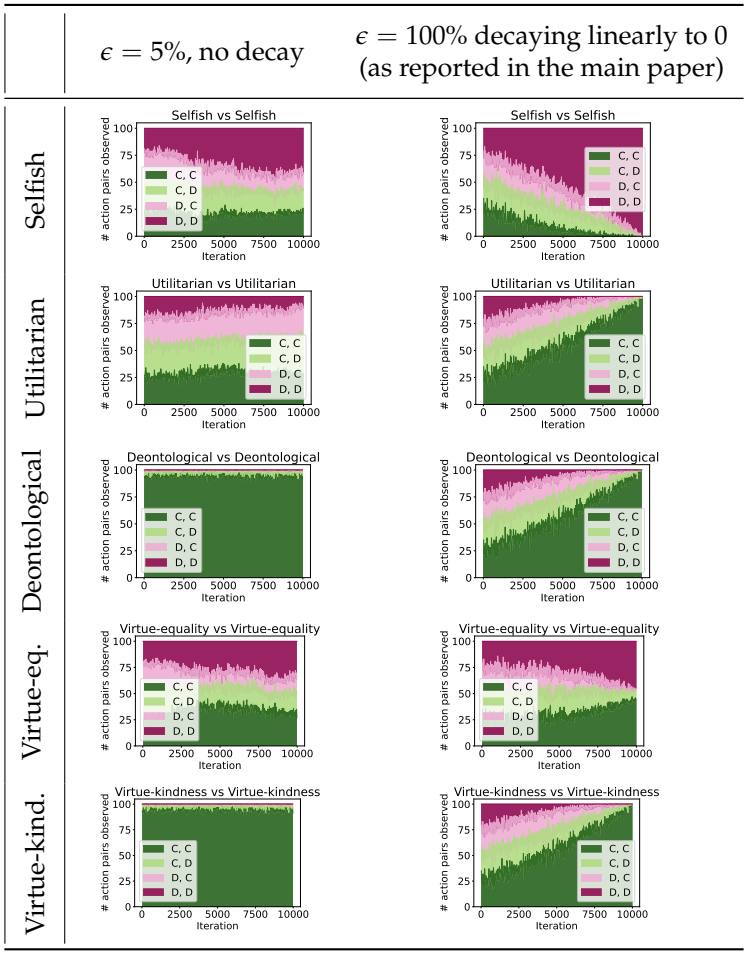


Figure 21: Iterated Prisoner's Dilemma game. The simultaneous action plots illustrate the impact of exploration on the learning of moral agents. For simplicity, we show each moral agent learning against its own kind only, and compare learning with a smaller exploration rate (left, $\epsilon = 5\%$) versus the large exploration rate (right, $\epsilon = 100\%$ decaying to 0), as reported in the paper.

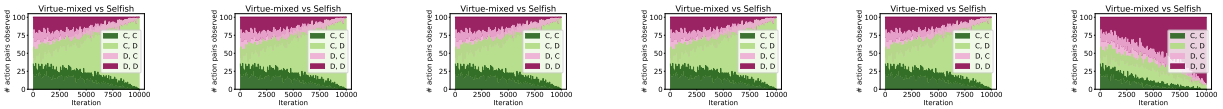
	vs QLS; $\beta = 0$ (fully 'kind')	vs QLS; $\beta = 0.2$	vs QLS; $\beta = 0.4$	vs QLS; $\beta = 0.6$	vs QLS; $\beta = 0.8$	vs QLS; $\beta = 1.0$ (fully 'equal')
Virtue-mixed						

Figure 22: Iterated Prisoner's Dilemma. The actions displayed by *Virtue-mixed* agent defined with different weights $\beta \in (0, 0.2, 0.4, 0.6, 0.8, 1)$ against a *Selfish* opponent.

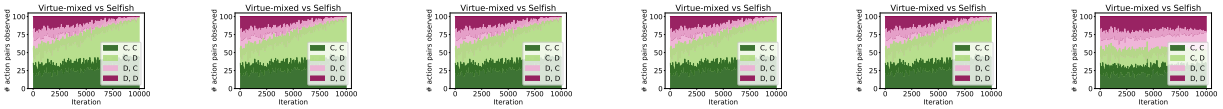
	vs QLS; $\beta = 0$ (fully 'kind')	vs QLS; $\beta = 0.2$	vs QLS; $\beta = 0.4$	vs QLS; $\beta = 0.6$	vs QLS; $\beta = 0.8$	vs QLS; $\beta = 1.0$ (fully 'equal')
Virtue-mixed						

Figure 23: Iterated Volunteer's Dilemma. The actions displayed by *Virtue-mixed* agent defined with different weights $\beta \in (0, 0.2, 0.4, 0.6, 0.8, 1)$ against a *Selfish* opponent.

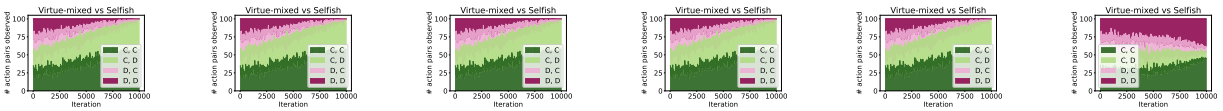
	vs QLS; $\beta = 0$ (fully 'kind')	vs QLS; $\beta = 0.2$	vs QLS; $\beta = 0.4$	vs QLS; $\beta = 0.6$	vs QLS; $\beta = 0.8$	vs QLS; $\beta = 1.0$ (fully 'equal')
Virtue-mixed						

Figure 24: Iterated Stag Hunt. The actions displayed by *Virtue-mixed* agent defined with different weights $\beta \in (0, 0.2, 0.4, 0.6, 0.8, 1)$ against an *Selfish* opponent.

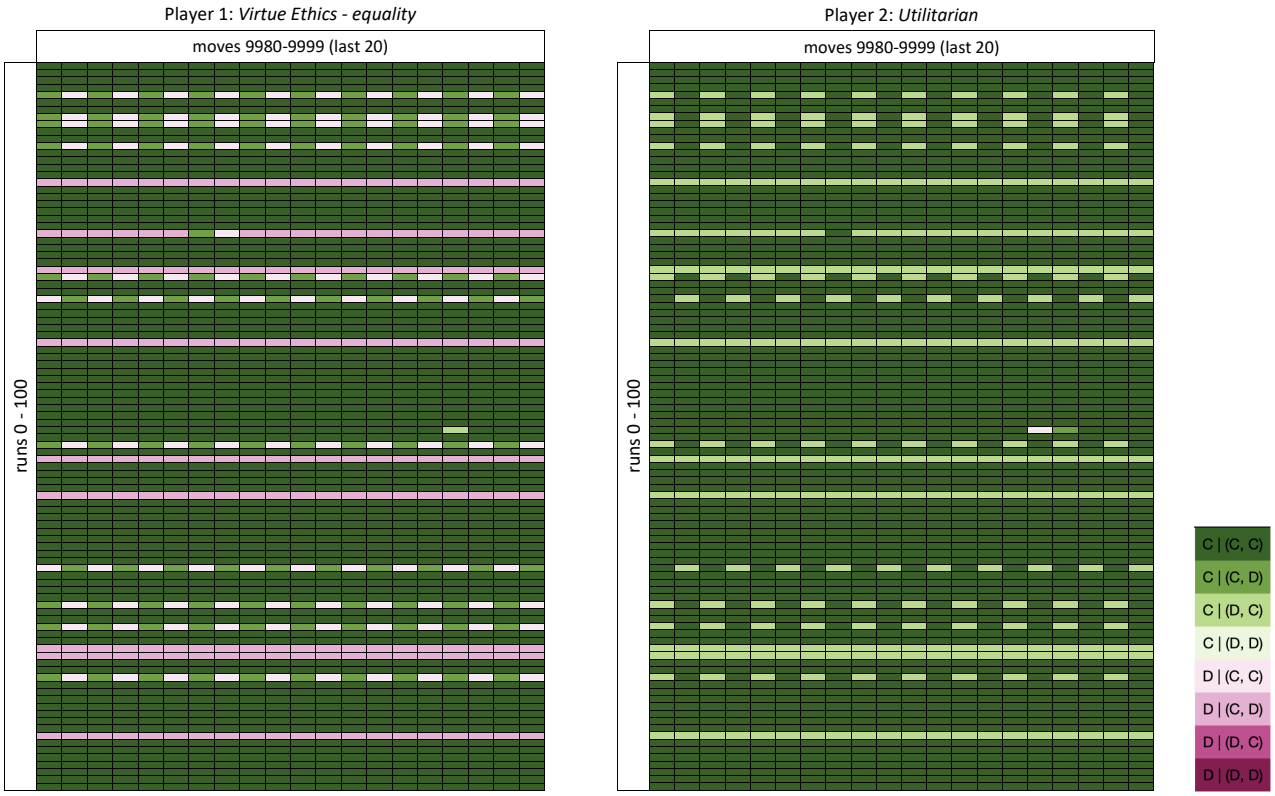


Figure 25: Iterated Prisoner's Dilemma. The last 20 actions played by the *Virtue-equality* agent (left) and the *Utilitarian* opponent (right), compared across the 100 runs. For each agent, we display each action given the state that the agent observed (see legend on the right). Each row represents a single run, and the 20 columns represent the last 20 consecutive moves observed (out of 10000).