Reinforcement Learning Agents for Interacting with Humans

Ido Shapira
Computer Science Department, Ariel University

Amos Azaria
Computer Science Department, Ariel University

Abstract

We tackle the problem of an agent interacting with humans in a general-sum environment, i.e., a non-zero sum, non-fully cooperative setting, where the agent's goal is to increase its own utility. We show that when data is limited, building an accurate human model is very challenging, and that a reinforcement learning agent, which is based on this data, does not perform well in practice. Therefore, we propose that the agent should try maximizing a linear combination of the human's utility and its own utility rather than simply trying to maximize only its own utility. We provide a formula to compute what we believe to be the optimal trade-off for the ratio between the human's and the agent's utility when attempting to maximize the agent's utility. We show the performance of our proposed method in two different domains. That is, our proposed agent not only maximizes the social welfare of both the human and the autonomous agent, but performs significantly better than agents not accounting for the human's utility function in terms of the agent's own utility.

Keywords: Human modeling, Human-agent, interaction, Reinforcement Learning.

Introduction

Autonomous agents interacting with humans are becoming ubiquitous. They are present in smart home environments, such as Alexa, Cortana, and Google Assistant, on the internet, as a form of chatbots or assisting bots, and in the physical world, such as robotic vacuum cleaners and mopers. Clearly, the presence of such agents will grow significantly in the years to come, including new areas, in which autonomous agents are only beginning to enter, such as autonomous vehicles, drones and other autonomous robots. Autonomous agents also interact with humans in competitive game environments, such as Chess, Go, Dota, and Starcraft (Skinner & Walmsley, 2019; Silver et al., 2016; Hsu, Campbell, & Hoane Jr, 1995).

For autonomous agents to proficiently interact with humans they must model human behavior. Such agents cannot rely on game theory or platforms assuming that humans are perfectly rational for composing a human model, as research into humans' behavior has found that people often deviate from what is thought to be rational behavior. People are affected by a variety of (sometimes conflicting) factors: a lack of knowledge of one's own preferences, the effects of the task complexity, framing effects, the interplay between emotion and cognition, the problem of self-control, the value of anticipation, future discounting, anchoring and many other effects (Tversky & Kahneman, 1981; Loewenstein, 2000; Ariely, Loewenstein, & Prelec, 2003; Camerer, 2003). Therefore, algorithmic approaches that use a pure theoretically analytic objective often perform poorly with real humans (Peled, Gal, & Kraus, 2011; Nay & Vorobeychik, 2016; Azaria, Aumann, & Kraus, 2012).

Unfortunately, rather than accounting for human utility, a widely common assumption, made by many works developing agents interacting with humans, is that an environment is either fully cooperative, and the agent's goal is identical to that of the human's or fully competitive, i.e., a zero-sum game. This assumption allows agent developers to ignore the utility function of the human, and concentrate only on maximizing the agent's utility function. While zero-sum and fully cooperative games are simpler to analyze, it is nearly impossible to find *any* real-life interaction between a group of humans that adheres to one of these assumptions. For example, even when two human players play a zero-sum board, card or sport game, their goal is usually to enjoy the interaction. Albeit, the winner might enjoy the overall experience better. If two human players were to play a true zero-sum game, one would need to kill the other in order to gain maximal utility. Furthermore, even a life-or-death gunfight pistol duel cannot be seen as a zero-sum game, as each of the players may attempt to not show, escape, give-up or only injure the opponent—all actions leading to outcomes that are not directly opposite. Similarly, fully cooperative games are not present in real-life either. Consider a married couple; they surely have some shared goals, such as raising their children and living in a place they are happy to be in. However, each individual has goals she cares more about than others, such as success at her own career or social life. If it were a fully cooperative game, couples would never argue and fight, and, clearly, there would be no divorcements. Researchers collaborating have a shared goal, but also have their own direction they would like the research to flow. They have the tasks they would like to contribute more and those they would like to contribute less to. Even when considering a limited task given to a group of people, some might want to be more dominant and instruct the others, which may, in turn, want to come up with a solution themselves.

Currently a common approach for developing an agent able to interact with humans in a general game (which is neither zero-sum nor fully cooperative) is by encapsulating

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

human behavior into a fixed model and ignoring the user's utility. This is usually performed by using machine learning techniques on a dataset, and possibly by also building upon psychological factors and human decision-making theory. The human behaviour model is then used by a planner to interact with humans (Gal & Pfeffer, 2007; Hindriks & Tykhonov, 2008; Subrahmanian, 2000; Rosenfeld & Kraus, 2011; Rosenfeld, Azaria, Kraus, Goldman, & Tsimhoni, 2015; Bitan, Gal, Kraus, Dokow, & Azaria, 2013). Other approaches, such as model free reinforcement learning, treat the human as a part of the environment and merely learn which actions the agent should take at which situations, in order to maximize its own reward (Carroll et al., 2019). However, an agent observing a sequence of actions performed by a human must ask why these actions were performed. Predicting future actions without accounting for the human's utility, is like predicting future words of an answer without accounting for the question being asked.

Furthermore, composing a human behavior model based on a relatively small data-set may be inaccurate, as people are many times unpredictable and different humans tend to behave differently from one another, despite a game being relatively simple (Shvartzon et al., 2016; Azaria, Richardson, & Rosenfeld, 2016). Therefore, we introduce a novel general method for solving the non-perfect human behavior model problem. We propose to maximize a linear combination of the agent's outcome and the human's outcome. We expect that optimizing toward a linear combination will be beneficial for the agent, since the humans are likely to try and optimize their own utility function, so they are likely to deviate from the human model in a way that will indeed maximize their utility function. By optimizing toward a linear combination, the agent acts as if it already accounts for these deviations and is therefore more likely to adapt to them. Moreover, we believe that referring to the human's reward will lead to cooperation that may be beneficial to the agent. That is, the human may act be more collaborative if the agent also tries to maximize the human's utility. Conversely, if the agent act completely selfishly, it is likely that the human will cooperate less and might take revenge on the agent, even if the human will lose from such actions. We provide a formula for determining the proposed linear combination, which is based on the similarity of the agents' utility functions and the accuracy of the human model. Namely, we introduce our Socially Aware Reinforcement Learning agent (SARL), an agent that attempts to maximize the linear combination of the two utility functions, using our proposed formula. We evaluate SARL in the following two domains: the single track road game introduced by (Shapira & Azaria, 2021), in which two agents are placed at different sides of a grid and must exchange places without colliding with each-other (see Figure 1). The second domain is a cleaning game, in which two agents are required to clean dirt, but each agent encounters a larger cost for moving than for remaining in its location (see Figure 2). We show that SARL significantly outperforms all other baselines in both domains when interacting with humans, in terms of the agent's final outcome.

Figure 1: The initial state of the single road game board. The red circle is controlled by the human player and the blue circle is controlled by the autonomous agent. Both players must reach the opposite side of the board without colliding. The players may travel freely on the upper row, but they cannot advance when located on the lower row.

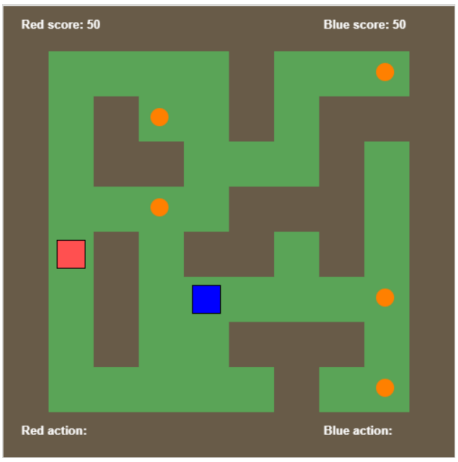

Figure 2: A screen-shot from the cleaning game board. The red square is controlled by the human player and the blue square is controlled by the autonomous agent.

To summarize, the main contribution of this paper is to present SARL, a socially aware reinforcement learner, that uses a linear combination of the rewards of both agents. We provide a formula for finding the parameter to be used in this linear combination. Finally, we show that SARL significantly outperforms all other baselines in two different domains.

Related Work

The first domain in which we test SARL is the single track road game. This game is somewhat similar to the repeated chicken-game, as introduced by Elhenawy et al. (2015). They introduce a real time game theory-based algorithm for controlling autonomous vehicle movements at uncontrolled intersections. They assume that all vehicles communicate to a central management center in the intersection to report their speed, location and direction. The intersection management center uses the information from all vehicles approaching the intersection and decides which action each vehicle will take. They further assume that vehicles obey the *Nash-equilibrium* solution of the game and will take the action received from the management center. Unfortunately, these assumptions are very strong and cannot be applied to our setting.

Camara et al. (2018) suggest a more realistic game-theory model based on the *sequential chicken-game*. The model assumes both agents share the same parameters U_{crash} and U_{time}, both know this is the case, and both play optimally from their state. It assumes that no lateral motion is permitted, and that there is no communication between the agents other than seeing each other's positions. The sequential chicken-game can be viewed as a sequence of one-shot (sub-)games, which can be solved similarly. The sub-game at time t can be written as a standard game theory matrix, which can be solved using recursion, and equilibrium selection to give values and optimal strategies at every state. While they handle the case of a junction by finding a Nash equilibrium and assuming that humans obey it, we provide a novel solution that does not require assumptions about humans and Nash equilibria.

There have been several previous works attempting to model human behavior in normal form games (Wright & Leyton-Brown, 2010, 2014). Wright and Leyton-Brown (2010) collected the results of multiple experiments from normal form games studied in the literature, and showed how the human action distribution can be modeled with high accuracy. However, our problem is clearly more complex and cannot be modeled as a simple normal form game.

Azaria el al. (2012; 2016) introduce SAP, a social agent for advice provision. They show that humans tend to ignore advice provided by a selfish agent. Therefore, they suggest using some linear combination of the user's and the agent's preferences. The exact ratio is determined by simulating human behavior and selecting the ratio that achieves the highest performance for the agent in simulation. Therefore, both SAP and our work attempt to maximize agent performance and consider a linear combination of both the user and the agent, however, the environment and settings are completely different, as SAP is an agent for advice provision, and we use a grid environment. In addition, the purpose of the linear combination used by SAP is to address the issue of human trust, while in our work, it is used to mitigate the uncertainty we have in our human model. Furthermore, we propose a formula for obtaining our proposed ratio, rather than running a simulation for obtaining that value.

There is a body of work on a group of complex games, which are non-zero sum nor fully cooperative, named social dilemmas. These games are a form of generalization of the iterated prisoner's dilemma, in which each agent may either decide to cooperate or defect (Sandholm & Crites, 1996; Wang, Zhou, Lien, Zheng, & Xu, 2016). Since this is an iterative process, the agents learn to cooperate (Vassiliades & Christodoulou, 2010). Most work in this field considers autonomous agents only, and does not consider humans (Jaques et al., 2019). One notable exception, though in the context of negotiation, is the colored trails game, which was developed to allow humans and agents to interact with each-other (Grosz, Kraus, Talman, Stossel, & Havlin, 2004).

Jaqus at el. (2018), propose a unified method for achieving both coordination and communication in MARL by giving agents an intrinsic reward for having a causal influence on other agents' actions. At each timestamp, the agent simulates alternate actions that it could have taken and runs a model of the other agents to see how influential each of its actions can be. Actions that lead to bigger changes in other agents' behavior are considered influential and are rewarded.

Gal et al. (2004) present an approach to modeling human behavior in one-shot games. The model predicts how a human player is likely to react to different actions of another player, and these predictions are used to determine the best possible strategy for that player. The authors found six possible influence features that they claim to reflect the human decisions in the game discussed.

Cooper et al. (2019), examine the idea of using a strategy that adaptively discourages antisocial behavior. Their proposed strategy has the overall structure of the folk theorem" of repeated games-stabilize but with punishment strategy that only restricts the opponent's utility to some safe target level while maximizing the utility of the agent. Clearly their proposed strategy cannot be used in our game since our game is not a repeated game.

Joseph et al. (2016) investigate the behavior of single-agent Q-learning in multi-agent environments. Their goal is to learn how the agent can be more cooperative without sacrificing their own individual rewards. This is quite different from our assumption that the agent attempts to maximize its own outcome.

Socially Aware Reinforcement Learning (SARL)

We introduce the Socially Aware Reinforcement Learning agent (SARL). SARL is a reinforcement learning agent that, similarly to common practice, treats the human as a part of the environment. It relies on a human model that is trained on data gathered from a human interacting with other agents. However, since the human model is likely to be inaccurate, instead of trying to maximize the agent's outcome directly, SARL uses a linear combination of its own outcome and the human's outcome. It is important to note that SARL is still selfish; it considers the human's outcome only because this is its way to maximize its own outcome. It is interesting to note that it has been shown in the field of psychology that people who consider other people's goals and show empathy, feel better with themselves and are more likely to reach their own goals (Carey, Tai, & Griffiths, 2021). Furthermore, reciprocation and cooperation may result in the human returning a favor. To that end, we define the parameter β, a value between 0 and 1, that quantifies the degree to which the agent considers its own outcome and the human's outcome. Namely, the agent, A, instead of optimizing towards $u(A)$, optimizes towards $\beta u(A) + (1-\beta)u(B)$. We note that when $\beta = 1$ the agent optimizes towards its own outcome. A β value of 0.5 entails that the agent tries to optimize the social outcome (i.e., $0.5u(A) + 0.5u(B)$, which is identical to optimizing simply towards $u(A) + u(B)$), and when $\beta = 0$ the

agent only considers the human's utility function.

β formula for SARL

Next, we make several assumptions and derive the optimal β value for SARL under these assumptions. Let μ be the accuracy of the human model (for a single step). Let H be the maximum expected accumulated return under the optimal policy (accounting for human's actions), and L a low expected accumulated return.

We further assume that, if optimizing toward the human's utility function, the optimization will work well (perfectly), as the human will adapt to all changes and assist in pursuing her own utility function. That is, if we optimize toward the human's utility, it is likely that the human will deviate in ways that will improve her own utility, so we are at least as likely to obtain the value that the agent expects in terms of human utility. Let ρ be the correlation between the accumulated rewards obtained by agents and humans playing the game (as we have previously defined). The following is our first attempt for formulating the true expected reward for the agent, v. For formula 1 we use the assumptions above and further assume that for the portion of which the agent optimizes toward its own utility, it will receive a value proportionate to the accuracy of the human model, and for the portion of which the agent optimizes toward the human's utility it will receive a value according to the human's utility and its correlation to the agent's utility.

$$v = \beta(\mu \cdot H + (1-\mu)L) + (1-\beta)(L + \rho \cdot (H-L)) \quad (1)$$

The derivative with respect to β is a constant; therefore, depending on μ, ρ, L,and H, one should either set β to its maximal value, 1.0, or its minimum value, 0. However, one cannot assume that the accuracy of the human model, μ, will persist also when the human model is used for optimization. We therefore assume that the further away from the human's utility the agent optimizes toward, the more inaccurate the model becomes. Therefore, our next attempt for formulating the true expected reward for the agent is:

$$v = \beta((1-\beta)\mu \cdot H + (1-(1-\beta)\mu)L) + (1-\beta)(L + \rho \cdot (H-L)) \quad (2)$$

Formula 2 entails that when using a selfish agent (with $\beta = 1$), the accuracy of the human model drops to 0 and the agent will obtain a value of L. A more realistic approach may assume that the human model accuracy halves rather than dropping down to 0. This yields our final formula:

$$v = \beta((1-\frac{\beta}{2})\mu \cdot H + (1-(1-\frac{\beta}{2})\mu)L) + (1-\beta)(L + \rho \cdot (H-L)) \quad (3)$$

If we differentiate Equation 3 with respect to β and set to 0 we obtain that the optimal β is given by:

$$\beta_{opt} = 1 - \frac{\rho}{\mu} \quad (4)$$

Indeed, we use the β obtained in Equation 4 as our β value for SARL, and demonstrate its performance in the two domains tested in this paper.

Experimental Design

We evaluate SARL's performance in the following two domains. In the single track road problem there are two vehicles in opposite directions must cross a narrow road, which is not wide enough to allow both vehicles to pass at the same time. Therefore, one vehicle must deter from to the other and let the other vehicle cross. We model the single track road problem as a sequential two player game on a two row grid (see Figure 1). The upper row represents a road that allows both players to advance. However, the lower row can only be used for allowing the other player to pass, as the players cannot advance when placed in the lower row. The reward function is defined as follows: a collusion ends the game and each agent encounters a loss of 100 points. An agent that arrives at its destination entails a reward of 30 points. Any agent that did not arrive at its destination and did not collide, obtains a penalty of 1 for continuing the game.

In the cleaning game, the two players are placed on a 10×10 grid board with 5 pieces of dirt that need to be cleaned (see figure 2). Both players can move only on the green area. The actions available for each player are: stay, left, up, down and right. Both agents begin with 50 points. Cleaning a piece of dirt does not cost or provide any points (until all dirt is clean). A move costs 5 points and remaining in place costs 1 point. Once all dirt is cleaned both players receive 100 points (regardless of how many dirt pieces each player has cleaned).

We recruited participants from Mechanical Turk (Paolacci, Chandler, & Ipeirotis, 2010) to play the two domains, 470 participants for playing the single road game and 412 for playing 3 variations of the cleaning game.

The participants first read the game instructions and were then required to answer three short and simple questions, to ensure that they had read and understood the instructions. The participants then played the game only once. Upon completion, the participants provided demographic information. In addition, following (Shapira & Azaria, 2021), each participant was asked to state how much they agreed with the following statements: 1. The agent played aggressively. 2. The agent played generously. 3. The agent played wisely. 4. The agent was predictable. 5. I felt the agent was a computer. Similarly, the participants in the cleaning game were also asked to state how much they agreed with five statements; however, the first two statements were slightly modified to match the cleaning game and were replaced by: 1. The agent played selfishly. 2. The agent was collaborative. We used a seven point Likert-like scale (Joshi, Kale, Chandel, & Pal, 2015) ranging from strongly disagree (1) to strongly agree (7).

For the single track road problem We reflect the results presented in (Shapira & Azaria, 2021) for the five baseline agents. In addition we run the following agents: 1. *Equal Social VI*: uses value iteration and the velocity human model to maximize the sum of the agent's and the human's utilities (i.e., $\beta = 0.5$). 2. *SARL*: uses value iteration and the velocity human model to maximize the linear combination of the

agent's and the human's utility computing β by Equation 4.

For the cleaning game we run the following agents: 1. *Selfish:* an agent that stays in place the entire game (and does not assist with cleaning the dirt). 2. *Closest:* an agent that always moves to the closest dirt. 3. *Farthest:* an agent that moves to the farthest dirt that will still come before the other player, in case there is no dirt like that, it stays in place. 4. *TSP:* an agent that moves by the solution of the TSP problem (Laporte, 1992). 5. *Random:* an agent that moves randomly. 6. *DDQN:* a DDQN agent that is trained on a custom openAI gym environment that we have created. A state of the environment is composed of an RGB image of the board, the dirt positions and the position of the two players. In order to incorporate movement, each move, the previous position of each of the players is added to the current state, but with an exponential discount rate of 0.9. The human model is trained based on the human behavior when playing against the baseline agents (items 1 to 5 in this list). The human model is perceived as a part of the environment. The human model is composed of a neural network with the input being a state and the output being a distribution over the human actions. The neural network consists of three convolutional layers with 8 kernels of size 4×4, 16 kernels of size 4×4, and 16 kernels of size 3×3. Followed by a max pooling layer with a size of 2×2, and a final convolutional layer with 8 kernels of size 3×3. Every convolutional layer uses a padding of 'same' and a ReLU activation function. Finally, there are two feed forward layers with sizes of 200 and 32 neurons, with ReLU activation. We use 80% of the data for training and 20% for validation. The accuracy of the human model was between 0.81 and 0.867 on the validation set. 7. *SARL:* based on the DDQN agent and uses the human model, but considers the other player's outcome by a linear combination of the two outcomes computed using Equation 4.

Results

In this section we present a comparison of all agents mentioned above and show that SARL significantly outperforms all other agents. We note that the baselines agents are used not only for comparison but also as our method of gathering data for composing the human model. The main competitor for SARL is DQN, which also uses the human model and thus, our analyses are focused mostly on comparing the two.

Result for the single track road game

The agent's score is calculated by averaging all its scores in each game it plays. Table 1 presents the performance of each of the agents along with the performance of the humans playing against them. As depicted by table 1, SARL significantly outperforms all other agents ($p < 0.01$) in terms of the agent's performance, and is the only agent that achieved a positive average reward. In addition, SARL also significantly outperforms all other agents ($p < 0.01$) in terms of social welfare. Surprisingly, the humans interacting with SARL performed better than the humans interacting with all other baselines; however, as will be shown, this result does not carry out to the second domain. Indeed, the β value for SARL in this game was 0.13, i.e., due to the high correlation between the performance of both players, and the relatively low accuracy of the human model, SARL mostly tried to maximize the human's performance, and was 87% altruistic and only 13% selfish. As we later show, in the second domain the correlation between the performance of both players is much lower, and the human model's accuracy is higher, resulting in much higher values for β.

Table 1: A comparison between the performance of each of the agents along with the human player who played against each of them.

	Avg. agent's score	**Avg. human's score**	**Avg. social welfare**
Careful	-2.29	-0.86	-3.15
Aggressive	-16.27	-18.40	-34.67
Semi-aggressive	-60.97	-62.11	-123.08
Random	-59.40	-57.62	-117.02
Velocity VI	-5.33	-6.03	-11.36
Eq. Social VI	-2.35	-4.09	-6.44
SARL	**15.87**	**17.12**	**32.99**

In addition, we tested the performance of a velocity value iteration agent with $\beta = 0$. That is, an agent that only considers the human reward. Interestingly, such an agent simply moves down and remains there forever, so that it does not disturb the human player. Unfortunately, such an agent achieves a final outcome of $-\infty$ (or $-\frac{1}{1-\gamma}$) because it can never reach its destination, since when the human's reaches her goal, the agent is directly beneath her.

We now turn to analyze the survey results for each agent (see Table 2). Each value in the table is the average of all scores of the measured values: Aggressively, Computer, Generously, Wisely and Predictable. Note that the lower the 'Aggressively' and 'Computer' parameters, the better the performance. On the other hand, the higher the 'Generously', 'Wisely' and 'Predictable' parameters, the better the performance. By Table 2, SARL obtained the best results compared to the other agents among all parameters except to 'Predictable'. These results entail that SARL

Table 2: Survey results of all agents for the single track road game.

	aggress.	**comp.**	**gen.**	**wise**	**pred.**
Careful	3.94	5.70	4.23	4.92	4.28
Aggressive	5.04	5.83	3.28	4.59	4.97
Semi-agg.	4.57	5.73	3.21	4.33	4.52
Random	3.51	5.64	4.01	3.72	3.57
Velocity VI	4.82	6.01	4.20	4.72	4.76
Social VI	4.78	5.60	3.69	4.92	**4.98**
SARL	**3.30**	**5.58**	**5.14**	**5.01**	4.00

demonstrates a clear improvement over all other agents.

Results for the cleaning game

As for the cleaning game, we ran all agents on three different board maps. The results reported in this section are averaged over the three board maps. We begin by comparing the average performance of each of the agents.

Table 3: A comparison between the performance of each of the agents along with the human player who played against each of them.

	Avg. agent's score	**Avg. human's score**	**Avg. social welfare**
TSP	0.83	**0.86**	1.69
Closest	0.69	0.78	1.47
Farthest	0.5	0.53	1.03
Selfish	1.11	0.11	1.22
Random	0.24	0.26	0.6
DDQN	1.1	0.11	1.21
SARL	**1.19**	0.64	**1.83**

As shown in Table 3, SARL significantly outperforms all other agents ($p < 0.01$) in terms of its own utility. In addition, and similarly to the single road problem, SARL significantly outperforms all other agents ($p < 0.01$) in terms of social welfare. However, humans interacting with TSP resulted in the highest performance. This is not surprising, as SARL considers the human's utility in order to maximize its own utility, and increasing the human's utility is only a side-effect.

We now turn to analyze the survey results in the cleaning game as they appear in Table 4. Each value in the table is the average of all the scores of the measured values: Selfishly, Computer, Collaborator, Wisely and Predictable. Note that the lower the 'selfishly' and 'computer' parameters, the better the performance. On the other hand, the higher the 'collaborator', 'wisely' and 'predictable' parameters, the better the performance. As can be seen in Table 4, SARL compared to the DDQN agent, obtained better results even in terms of courteous and generous. These results entail that SARL demonstrates a clear improvement compared to the DDQN agent. The β values for SARL in the three games are: $0.419, 0.74$, and 0.615 respectively. We noticed that the participants in the first cleaning game were the most satisfied with SARL's behavior (compared to the second and third game), i.e., SARL was rated as collaborative and wise. As expected, in the second game they were the least satisfied with the SARL's, since the β value was the highest.

Table 4: Survey results of all agents for the cleaning game.

	selfish	**comp.**	**coll.**	**wise**	**pred.**
TSP	**3.07**	5.54	**5.26**	**5.42**	**5.12**
Closest	3.52	6	4.81	4.83	5.01
Farthest	4.59	**5.32**	3.41	3.66	4.37
Selfish	6.02	5.72	2.42	3.74	5.06
Random	5.7	5.83	3.08	3.6	4.69
DDQN	6.13	5.37	2.95	3.78	5.27
SARL	5.37	5.62	3.76	4.68	4.93

Next, we evaluate the average number of times the human players decided to remain in place and not help the agent (encountering a lower cost). Remaining in place may be either as an act of revenge against the other agent, who the human player believes to not assist enough, or as an attempt to work less and have the other agent work harder. We noticed that the participants were the most vindictive toward the selfish agent, with an average of 9.1 stays per game. Similarly, the participants performed 8.8 stays per game when playing with the DDQN agent. The participants also performed 'stay' actions when playing against the closest agent (3.1), the TSP agent (2.6), and the farthest agent (2.3), as they probably noticed that even if the participants do not help, the agents will continue working and completing the task. We note that the participants performed many 'stay' actions when playing against the random agent (4.9); this might be because they did not really understand what the agent was doing. Interestingly, the participants performed the least 'stay' actions when playing with SARL (1.21). This indicates that SARL achieved a high level of collaboration with the participants. All differences in the behavior and performance between different genders and level of education, were found to be non-statistically significant.

Conclusions and Future Work

In this paper we present SARL. We showed that when data is limited, building an accurate human model is very challenging, and that a reinforcement learning agent, which was based on this data, did not perform well in practice. However, we showed that a social agent, i.e., an agent that tried to maximize a linear combination of the human's utility and its own utility, achieved a high score, and significantly outperformed other agents, including an agent that simply tried to maximize only its own utility. We provided a formula to compute what we believe to be a good choice for the β parameter, i.e., the ratio between the human's and the agent's utility when attempting to maximize the agent's utility. In addition, we showed that the social welfare of both of the agents was highest when interacting with SARL. In future work we intend to show that SARL performs well also when considering other, possibly very different, settings. One option for such a setting is a setting with a continuous state space as well as a continuous action space. Another direction for future work is to focus on situations in which the human reward function is not available apriori. Such a situation would challenge the use of SARL, as it uses the human reward function for computing its objective function. One appealing option may be to use inverse reinforcement learning (Ng, Russell, et al., 2000) to first learn the human's reward function, and then, to use this function to compute the optimal policy for SARL.

Acknowledgment

This research was supported in part by the Ministry of Science, Technology & Space, Israel.

References

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106.

Austerweil, J. L., Brawner, S., Greenwald, A., Hilliard, E., Ho, M., Littman, M. L., ... Trimbach, C. (2016). How other-regarding preferences can promote cooperation in non-zero-sum grid games.

Azaria, A., Aumann, Y., & Kraus, S. (2012). Automated strategies for determining rewards for human work. In *Twenty-sixth aaai conference on artificial intelligence.*

Azaria, A., Gal, Y., Kraus, S., & Goldman, C. V. (2016). Strategic advice provision in repeated human-agent interactions. *Autonomous Agents and Multi-Agent Systems*, *30*(1), 4–29.

Azaria, A., Rabinovich, Z., Kraus, S., Goldman, C., & Gal, Y. (2012). Strategic advice provision in repeated human-agent interactions. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 26).

Azaria, A., Richardson, A., & Rosenfeld, A. (2016). Autonomous agents and human cultures in the trust–revenge game. *Autonomous Agents and Multi-Agent Systems*, *30*(3), 486–505.

Bitan, M., Gal, Y., Kraus, S., Dokow, E., & Azaria, A. (2013). Social rankings in human-computer committees. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 27).

Camara, F., Romano, R., Markkula, G., Madigan, R., Merat, N., & Fox, C. (2018, 03). Empirical game theory of pedestrian interaction for autonomous vehicles..

Camerer, C. F. (2003). Behavioral Game Theory. Experiments in Strategic Interaction. In (pp. 43–118). Princeton University Press.

Carey, T. A., Tai, S. J., & Griffiths, R. (2021). *Deconstructing health inequity: A perceptual control theory perspective*. Springer Nature.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. *Advances in Neural Information Processing Systems*, *32*, 5174–5185.

Cooper, M., Lee, J. K., Beck, J., Fishman, J. D., Gillett, M., Papakipos, Z., ... Littman, M. L. (2019). Stackelberg punishment and bully-proofing autonomous vehicles. *CoRR*.

Elhenawy, M., Elbery, A., Hassan, A., & Rakha, H. (2015, 09). An intersection game-theory-based traffic control algorithm in a connected vehicle environment. In (p. 343-347). doi: 10.1109/ITSC.2015.65

Gal, Y., & Pfeffer, A. (2007). Modeling reciprocal behavior in human bilateral negotiation. In *Proceedings of the national conference on artificial intelligence* (Vol. 22, p. 815).

Gal, Y., Pfeffer, A., Marzo, F., & Grosz, B. (2004, 01). Learning social preferences in games.

Grosz, B., Kraus, S., Talman, S., Stossel, B., & Havlin, M. (2004). The influence of social dependencies on decision-making: Initial investigations with a new game.

Hindriks, K., & Tykhonov, D. (2008). Opponent modelling in automated multi-issue negotiation using bayesian learning. In *Aamas* (pp. 331–338).

Hsu, F.-h., Campbell, M. S., & Hoane Jr, A. J. (1995). Deep blue system overview. In *Proceedings of the 9th international conference on supercomputing* (pp. 240–244).

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., ... De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning* (pp. 3040–3049).

Jaques, N., Lazaridou, A., Hughes, E., Gülçehre, Ç., Ortega, P. A., Strouse, D., ... de Freitas, N. (2018). Intrinsic social motivation via causal influence in multi-agent RL. *CoRR*.

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, *7*(4), 396.

Laporte, G. (1992). The traveling salesman problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, *59*(2), 231-247.

Loewenstein, G. (2000). Willpower: A decision-theorist's perspective. *Law and Philosophy*, *19*, 51-76.

Nay, J. J., & Vorobeychik, Y. (2016). Predicting human cooperation. *PloS one*, *11*(5), e0155656.

Ng, A. Y., Russell, S. J., et al. (2000). Algorithms for inverse reinforcement learning. In *Icml* (Vol. 1, p. 2).

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411–419.

Peled, N., Gal, Y. K., & Kraus, S. (2011). A study of computational and human strategies in revelation games. In *Aamas* (pp. 345–352).

Rosenfeld, A., Azaria, A., Kraus, S., Goldman, C. V., & Tsimhoni, O. (2015). Adaptive advice in automobile climate control systems. In *Workshops at the twenty-ninth aaai conference on artificial intelligence.*

Rosenfeld, A., & Kraus, S. (2011). Using aspiration adaptation theory to improve learning. In *Aamas* (pp. 423–430).

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

Shapira, I., & Azaria, A. (2021). Autonomous agents for the single track road problem. In *2021 ieee 33rd international conference on tools with artificial intelligence (ictai)* (pp. 81–85).

Shvartzon, A., Azaria, A., Kraus, S., Goldman, C. V., Meyer, J., & Tsimhoni, O. (2016). Personalized alert agent for optimal user performance. In *Thirtieth aaai conference on artificial intelligence.*

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Skinner, G., & Walmsley, T. (2019). Artificial intelligence and deep learning in video games a brief review. In *2019 ieee 4th international conference on computer and communication systems (icccs)* (pp. 404–408).

Subrahmanian, V. S. (2000). *Heterogeneous agent systems*. MIT press.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458.

Vassiliades, V., & Christodoulou, C. (2010). Multiagent reinforcement learning in the iterated prisoner's dilemma: fast cooperation through evolved payoffs. In *The 2010 international joint conference on neural networks (ijcnn)* (pp. 1–8).

Wang, Z., Zhou, Y., Lien, J. W., Zheng, J., & Xu, B. (2016). Extortion can outperform generosity in the iterated prisoner's dilemma. *Nature communications*, *7*(1), 1–7.

Wright, J. R., & Leyton-Brown, K. (2010). Beyond equilibrium: Predicting human behavior in normal-form games. In *Twenty-fourth aaai conference on artificial intelligence.*

Wright, J. R., & Leyton-Brown, K. (2014). Level-0 meta-models for predicting human behavior in games. In *Proceedings of the fifteenth acm conference on economics and computation* (pp. 857–874).