Reinforcement Learning Agents and Human Interaction

Ido Shapira [a,*] and Amos Azaria [b]
[a] *Department of Computer Science, Ariel University, Israel*
E-mail: idos@ariel.ac.il
[b] *Department of Computer Science, Ariel University, Israel*
E-mail: amos.azaria@ariel.ac.il

Abstract. While autonomous agents that interact with humans are becoming more and more prominent, currently, state-of-the-art methods first compose a model of human behavior, which is based on observing human actions, and then optimize the agent's actions based on this model. However, such methods do not usually account for how the human will react to the agent's actions and thus, suffer an overestimation bias. Therefore, in this paper, we pursue solution concepts for autonomous agents interacting with humans that overcome the overestimation bias problem. We propose three solution concepts: a social optimization approach, empathy based learning, and pessimistic execution. In the social optimization approach, rather than only optimizing toward the agent's utility function, the agent optimizes toward a linear combination of the agent's and the human's utility functions. In the empathy based learning approach, the agent attempts to predict participants' attitudes toward several aspects of the agent. In the pessimistic execution approach, the agent composes two or more human models, samples from each of the models and assumes that the human takes the action that is least beneficial for the agent. Our results show that our methods outperform other baselines, in terms of agent's utility, in three different environments.

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

1. Introduction

Autonomous agents interacting with humans are ubiquitous. They are present in smart home environments, such as Alexa, Cortana, and Google Assistant, on the internet, as a form of chatbots, assisting bots, and large language models (e.g., ChatGPT [1]), and in the physical world, such as robotic vacuum cleaners and mopers. Clearly, the presence of such agents will continue to grow 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 [2–4].

Autonomous agents attempting to proficiently interact with humans 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 people often deviate from what is thought to be rational behavior. Namely, people are affected by a variety of 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 [5–8]. Therefore, algorithmic approaches that use a pure theoretically analytic objective often perform poorly with real humans [9–11]. Nevertheless, the concept of using a utility function to explain human behavior is very common and has been widely used in economics and psychology for centuries. Prior to the 1700s, it was common to use the expected *monetary gain* as the utility function, which people are assumed to maximize. The expected utility hypothesis, formulated by Daniel Bernoulli, states that people attempt to maximize the expected utility, rather than their expected monetary gain. The prospect theory and the cumulative

*Corresponding author. E-mail: idos@ariel.ac.il.

prospect theory, which are further refined to the expected utility hypothesis, awarded Daniel Kahneman a Nobel prize [12]. It is clearly a fundamental assumption that humans performing some task, are doing it in order to achieve some goal. Therefore, modeling human behavior must account for human goals.

A common approach for developing agents interacting with humans in a general game (which is neither zero-sum nor fully cooperative) is by encapsulating human behavior into a fixed model and ignoring the human'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 behavior model is then used by a planner to interact with humans [13–16]. Unfortunately, such approaches are prone to over-fit and suffer from the over-estimation bias [17, 18]. This is because human models are only accurate when all players behave as they did during the data-collection phase. Using the human model as a fixed model, does not account for the fact that human behavior also depends on the agent's behavior itself.

Therefore, we introduce three methods for overcoming the overestimation bias problem. Our first method attempts 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, therefore is, more likely to adapt to them. Moreover, we believe that referring to the human reward will lead to collaboration that may be beneficial to the agent. That is, the human may be more collaborative if the agent tries to maximize the human's utility as well. Conversely, if the agent acts 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 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.

Our second method, the Empathy based Reinforcement Learning (ERL) and third method, the Pessimistic Reinforcement Learning (PRL) are not based on the human reward function. Namely, the ERL agent attempts to enhance the human model by attempting to predict participant's rating of several aspects of the agent. That is, the model learns to predict not only the player's actions but also the attitude the human has toward the agent, which influences her actions. The PRL agent trains multiple human models. It then attempts to overcome the overestimation bias by sampling the human models and assuming that the human player will take the action that is least beneficial to the agent.

We evaluate our solution approaches in the following three domains: the single track road game introduced by [19], 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). The third domain is the stag-hunt game, in which two agents make a choice between a risky action (hunt the stag) and a safe action (forage for fruits). Foraging for fruits always yields a safe payoff while hunting yields a high payoff only if the other player also decides to hunt.

We show that SARL significantly outperforms all other baselines in those domains when interacting with humans, in terms of the agent's final outcome. Somewhat less surprisingly, humans interacting with SARL also achieve a high outcome. Therefore, SARL not only performs better with respect to its own outcome but also with respect to social welfare and collaboration. In addition, we show that the ERL and PRL agents, perform as well as the traditional RL agent, while improving the social welfare.

To summarize, the main contribution of this paper is threefold:

(1) We present SARL, a socially aware reinforcement learner, that uses a linear combination of the rewards of both agents.
(2) We provide a formula for finding the parameter to be used in this linear combination, and show that SARL significantly outperforms all other agents in three different domains.
(3) We present empathy based reinforcement learning and pessimistic reinforcement learning, two methods for agents interacting with humans, attempting to overcome the overestimation bias. We show the performance of these methods in different domains.

2. Related Work

See reviewer's note

In this paper we suggest three different methods for reinforcement learning agents interacting with humans and test those methods in three different environments. The first environment we use is the single-track road game. This game is somewhat similar to the repeated chicken game, as introduced by Elhenawy et al. [20]. 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. [21] 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.

The second environment used for testing our agents is the cleaning game. This game is somewhat similar to the cleanup game introduced by Jaques et al. [22]; however, the cleanup game focuses on limited communication between agents and is an attempt to reach coordination. In the cleaning game, the agents can see the entire board and whether each agent is working or resting. The third environment that we test our agents in is the stag-hunt game, which was introduced by Peysakhovich and Lerer [23]. In their work, they change the learning rule of a single agent to improve its outcomes in Stag Hunt environments that include other reactive learners. They also extend existing work on reward-shaping in multi-agent reinforcement learning and show that making an agent care also about the rewards of their partners can increase the probability that groups converge to good outcomes. Unfortunately, they do not test their approach in an environment with a human and an autonomous agent.

Sequential Social Dilemmas (SSDs) are firstly introduce by Leibo et al [24]. SSDs extended multi-agent games that have a payoff structure similar to that of Prisoner's Dilemma. That is, an individual agent can obtain higher reward by engaging in defecting, non-cooperative behavior (and thus is rationally motivated to defect), but the average payoff per agent will be higher if all agents cooperate. 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 [25, 26]. Since this is an iterative process, the agents learn to cooperate [27]. Most work in this field considers autonomous agents only and does not consider humans [28]. 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 [29].

Jaques et al. [22], propose a unified method for achieving both coordination and communication in Multi-Agent Reinforcement Learning (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 run 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.

Azaria el al. [30, 31] introduce SAP, a social agent for advice provision. They show that humans tend to ignore the 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 have been several previous works attempting to model human behavior in normal form games [32, 33]. Wright and Leyton-Brown [32] 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. Moreover, Gal et al. [34] 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. [35], 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 a 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. In addition, Joseph et al. [36] 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.

3. Reinforcement Learning Agents for Interacting with Humans

In this section we introduce our three approaches for developing reinforcement learning agents for interacting with humans, namely, Socially Aware Reinforcement Learning (SARL), Empathy Based Reinforcement Learning (ERL), and Pessimistic Reinforcement Learning (PRL). All three approaches rely on the common practice of treating the human as a part of the environment and relying on a human model that is trained on data gathered from a human interacting with other agents. However, they differ by how they approach the overestimation bias problem.

3.1. Socially Aware Reinforcement Learning (SARL)

Since the human model is likely to be inaccurate, instead of trying to maximize the agent's outcome directly, the Socially Aware Reinforcement Learning agent (SARL), uses a linear combination of its own outcome and the human's outcome. Namely, given a utility function for the human, U_{human}, a utility function for the agent U_{agent}, and a value of β controlling the "selfishness" of the agent, the agent optimizes toward:

$$U_{agent} = \beta \cdot U_{agent} + (1-\beta) \cdot U_{human}$$

.

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 about themselves and are more likely to reach their own goals [37]. Furthermore, reciprocation and cooperation may result in the human returning a favor.

3.1.1. 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 actions), and L is 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. 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.

$$v = \beta(\mu \cdot H + (1-\mu)L) + (1-\beta)(L + \rho \cdot (H-L)) \tag{1}$$

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 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:

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

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:

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

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

$$\beta_{opt} = 1 - \frac{\rho}{\mu} \tag{4}$$

In some scenarios, a β value less than 0.5 becomes non-plausible, since it means that the agent prefers that the human will obtain a point than itself. Therefore, in such scenarios, we normalize β between 0.5 and 1:

$$\beta_{opt} = 1 - 0.5 \cdot \frac{\rho}{\mu} \tag{5}$$

Indeed, we use the β obtained in Equation 4 as our β value for SARL in the single track-road problem and the cleaning game. As for the stag-Hunt game, we use the β obtained in Equation 5, and demonstrate its performance in the domains tested in this paper.

3.2. Empathy Based Reinforcement Learning (ERL)

SARL is an agent that attempts to consider the human's goals, thus it might need to gain access to a human's utility function or compose such a function. This may be a problem when interacting with humans in a real-world situation, as the utility function is not always available. There are several methods for obtaining such a utility function.

One option is by using inverse reinforcement learning, which by observing human actions, can elicit the underlying utility function [38]. A somewhat similar approach is to compose a neural architecture that learns both the utility function and the prediction of future human actions simultaneously. However, these approaches may be problematic, as our general hypothesis is that one should account for human utility function when composing a human model and not only rely on observing human actions.

Another approach is to allow humans to communicate their preferences to the agent and tell it what their utility function is. Clearly, it is not a simple task to create an interface that allows a human to communicate her complete utility function. Furthermore, people may not be fully aware of their utility function, may not be willing to fully disclose it, or may act strategically and not reveal their true utility function, so that the agent's actions will be more beneficial to them. A more suitable approach would be to have humans answer survey questions, similar to methods used for preference elicitation and in the field of ethics research.

The ERL agent attempts to understand the motive behind human actions and not only to predict them. To that end, each participant is requested to rate several aspects of the agent that they played with, representing the human's attitude toward the agent (see section 4 for the survey). The human model is then trained to predict not only the action taken by the human, but also the participant's attitude toward the agent. Therefore, instead of a single output layer, the ERL human model has two outputs layers: the first for the human action and the second for the participant's rating value. In addition, the participant's rating output serves as an input for the human action output. The loss is defined as follows:

$$loss = \alpha \cdot loss_a + (1 - \alpha) \cdot loss_r$$

Where $loss_a$ refers to the human action loss and $loss_r$ refers to the participant's rate. α was set to 0.8.

3.3. Pessimistic Reinforcement Learning (PRL)

In some situations it is not possible to obtain the human reward function required by SARL, nor is it possible to obtain participant's ratings, required by the ERL agent. Therefore, the PRL agent only requires as input observations of humans interacting in the environment. The PRL agent trains multiple human models, which differ merely by the original weights of a neural network, by their training data, or by using a dropout layer at inference. During traning of the reinforcement learning, each of the human models is sampled, and the agent assumes that the human will play that action that is least beneficial for the agent. This approach is anticipated to reduce the overestimation bias.

4. Experimental Design

In this section we introduce the three domains that we used in order to evaluate our agents performance alongside the baselines that used to gather human data and the RL agents we compared with.

4.1. The Single-Track Road problem

In the Single-Track Road problem, there are two vehicles in opposite directions that 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 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: 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. An agent that did not arrive at its destination and did not collide, obtains a penalty of 1 each step in the game.

We model the problem as an MDP in which the human's actions are modeled as a part of the environment. The model uses data from humans interacting with simple agents to determine the probability of the human taking each action at a given state. We define a state as a pair (i, j) in which i is a position of the autonomous agent, and j is a position of the human agent. We refer to this state representation as a state *without* velocity. We also use a more complex representation of a state by considering also the previous locations of both players. That is, a state is a tuple of two pairs $((i, j), (l, k))$, where the first pair, (i, j), is the current non-velocity state of the two agents, and the second pair, (l, k), is their previous non-velocity state. This representation is referred to as a state with velocity

4.1.1. Baselines agents

Introduce the five following baselines that used for collecting human data that first show by [19]:

(1) *Careful:* an agent that adheres to the strategy of agent B in Theorem . That is, it tries to moves left, but tries to avoid colliding with the other agent as well, so if moving left may risk colliding with the other agent it stays in place. If staying in place also risks colliding with the other agent, it moves down.
(2) *Aggressive:* an agent that adheres to the strategy of agent A in Theorem . That is, the agent always moves left.
(3) *Semi-aggressive:* an agent that moves left unless the other agent is already there, in which case it stays in place until the other agent moves out of its way.
(4) *Random:* an agent that moves randomly.

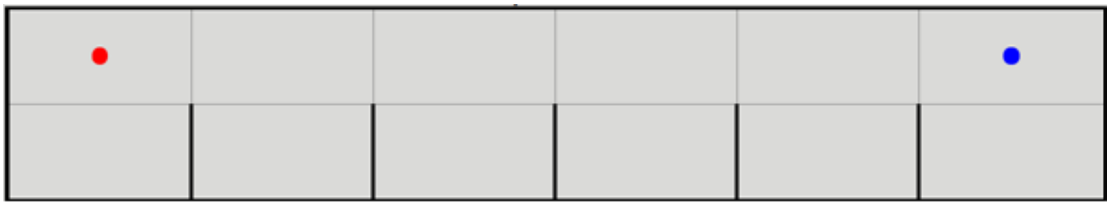

Fig. 1. The initial state of the single road game board. The red circle is controlled by the human player and the blue circle is controlled by the autonomous agent. Both players must reach the opposite side of the board without colliding. The players may travel freely on the upper row, but they cannot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

4.1.2. Competing agents

For competing with our agents we reflect the following RL agents that introduce in [19]:

(1) *Non-Velocity VI*: runs a value iteration on the MDP without velocity using the appropriate human model.
(2) *Velocity VI*: runs a value iteration on the MDP with velocity using the appropriate human model.

In addition, we ran the following RL 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.

4.2. The Cleaning 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 points. Once all dirt is cleaned both players receive 100 points (regardless of how many dirt pieces each player has cleaned).

4.2.1. Baselines agents

Introduce the five following baselines that used for collecting the human data:

(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 [39].
(5) *Random:* an agent that moves randomly.

4.2.2. Human model

An input for the human model is the state of the environment that composed of a 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 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 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 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.

Fig. 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.

4.2.3. Competing agents

Based on the human model introduce in 4.2.2 that being part of the environment we trained and evaluated the following RL agents:

(1) *DDQN:* a DDQN agent that is trained on a custom openAI gym environment that we have created.
(2) *ERL:* based on a DDQN agent but considers the participant's rating values alongside the human model training, as described in section 3.2.
(3) *PRL:* based on a DDQN agent but draws two actions from the human model and assumes that the next player's action will be least preferable for the agent 3.3.
(4) *SARL:* based on a 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.

4.3. The Stag-Hunt game

In the Stag-Hunt game, the two players are placed on a 5×5 grid board with 3 bushes randomly placed on the board and a stag (see figure 3). Both players can move freely inside the borders. The actions available for each player are: left, up, down, and right. The players play 60 moves. each time the players move, the stag chooses one of the three options uniformly: take a random move, move towards the nearest player, or stay in place. Forage for fruits provides 1 point. For hunting the stag both players must catch it together, in which case provides a high reward of 4 points. The goal is to get as high a cumulative reward as possible.

Fig. 3. A screen-shot from the stag hunt game. The red player is controlled by a human player and the blue player is controlled by an autonomous agent.

4.3.1. Baselines agents

Introduce the four following baselines that used for collecting the human data:

(1) *Closest:* an agent that ignores the other player and the stag and always goes to the closest bush.
(2) *Follow-stag:* an agent that always moves towards the stag.
(3) *Random:* an agent that moves randomly.

4.3.2. Human model

A state of the environment is a vector of the positions of both agents, bushes positions, and the stag position. An input for the human model composed of RGB image which was built based on the environment state. The human model is trained based on human behavior when playing against the baseline agents (items 1 to 3 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 human actions. The neural network consists of four convolutional layers with 4 kernels of size 3×3, 8 kernels of size 3×3, 16 kernels of size 3×3, and 8 kernels of size 2×2. Every convolutional layer uses padding of 'same' and a ReLU activation function. Finally, there is a feed-forward layer with sizes of 64, with ReLU activation. We use 80% of the data for training and 20% for validation. The accuracy of the human model was 0.724 on the validation set.

4.3.3. Competing agents

Based on the human model introduce in 4.3.2 that being part of the environment we trained and evaluated the following RL agents:

(1) *DDQN:* a DDQN agent that is trained on a custom openAI gym environment.

(2) *ERL:* based on the DDQN agent but considers the participant's rate alongside the human model training as described in section 3.2.
(3) *PRL:* based on the DDQN agent but assuming that the next action of the human player will be the most damaging for him as described in section 3.3.
(4) *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.

4.4. Experiments

We recruited participants from Mechanical Turk [40] to play the three domains, 470 participants for playing the single road game, 412 for playing 3 variations of the cleaning game, and 230 for playing the Stag-Hunt 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 [19], 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 and in the Stag-Hunt 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 [41] ranging from strongly disagree (1) to strongly agree (7).

5. Results

In this section, we present a comparison of all agents mentioned above and show that SARL significantly outperforms all other agents. The main competitor for SARL is DDQN, ERL, and PRL, which also uses a human model and thus, our analyses are focused mostly on comparing these four agents. All differences in the behavior and performance between different genders and levels of education were found to be non-statistically significant.

5.1. Results 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. For the five baseline agents we reflect the results presented in [19]. 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 next domains. 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.

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 reaches her goal, the agent is directly beneath her.

Table 1
A comparison between the performance of each of the agents along with the human player who played against each of them in the single track road game.

Agent	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**

5.1.1. The imperfection of the single-track road environment

We evaluate the prediction of the *policy evaluation* algorithm, using both forms of state representations (i.e., with and without velocity). Table 2 presents the prediction compared with the actual score of every agent. As can be seen

Table 2
The accuracy of the prediction of a policy evaluation algorithm using a model with velocity and a model without velocity.

Agent	True score	Prediction with velocity (error)	Prediction without velocity (error)
Careful	**-2.29**	-14.41 (12.12)	-4.86 (2.57)
Aggressive	-16.27	-6.21 (10.6)	1.14 (17.41)
Semi-aggressive	-60.97	-56.47 (4.5)	-47.81 (13.16)
Non-Velocity VI	-6.34	0.51 (6.85)	**13.63** (19.97)
Velocity VI	-5.33	**14.47** (20.02)	N/A
Social VI	-2.35	12.34 (14.69)	N/A
SARL	**15.87**	7.55 (8.32)	N/A

in the table, the prediction that uses a state representation with velocity outperforms the prediction that uses a state representation without velocity. However, both predictions performed badly, and imply that our human model is not accurate, as an accurate human model would have resulted in an accurate prediction. This demonstrates that it is not enough to rely on the dataset, and strengthens the need for the socially aware approach, which also considers the human rewards.

5.1.2. Survey results of the single track road game

We turn to analyze the survey results for each agent (see Table 3). 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',

Table 3
Survey results of all agents for the single track road game in the single track road game.

Agent	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

'Wisely' and 'Predictable' parameters, the better the performance. Table 3 shows that SARL, SARL obtained the best results compared to the other agents among all parameters except 'Predictable'. These results entail that SARL demonstrates a clear improvement over all other agents.

5.2. Results for the cleaning game

In 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 4

A comparison between the performance of each of the agents along with the human player who played against each of them for the cleaning game.

Agent	Avg. agent's score	Avg. human's score	Avg. social welfare
TSP	83	**86**	169
Closest	69	78	147
Farthest	50	53	103
Selfish	111	11	122
Random	24	26	50
DDQN	110	11	121
ERL	110	48	158
PRL	109	38	147
SARL	**119**	64	**183**

As shown in Table 4, 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 to maximize its own utility, and increasing the human's utility is only a side-effect. Also, we note that the traditional DDQN acted in many cases similar to the selfish baseline. It can explain why they both get very similar scores. In addition, while the ERL and the PRL do not assume any assumptions, they received a lower score than SARL when playing with the participants. Moreover, while the ERL and the PRL scores are similar to the traditional DDQN, their welfare is higher. This indicates that these agents cooperated more than the traditional DDQN. It is shown also in the survey results (See table 6).

5.2.1. The imperfection of the cleaning environment

We evaluate the prediction of the cleaning environment on the different agents, Table 5 presents the prediction compared with the actual score of every agent. As can be seen in the table, the prediction performed badly. The prediction indicates at the DDQN agent has the best policy, but actually in the experiment with the participants SARL got a better score which entails that the DDQN policy was not optimal. This demonstrates that it is not enough to rely on the dataset, and strengthens the need for the socially aware approach (or other approaches).

5.2.2. Survey results of the cleaning game

We analyze the survey results in the cleaning game as they appear in Table 6. 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 6, SARL compared to the DDQN agent, obtained better results even in terms of courtesy and generosity. 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. In addition, the ERL and the PRL agents cooperated more than SARL. The goal was to get much score as possible and not necessarily to cooperate or be generously but still, this result is worth mentioning.

Table 5
The accuracy of the prediction of all policy agents in the cleaning environment.

Agent	True score	Prediction (error)
TSP	83	59 (24)
Closest	69	64 (5)
Farthest	50	14 (36)
Selfish	111	102 (9)
Random	24	6 (18)
DDQN	110	**106** (4)
ERL	110	104 (6)
PRL	109	104 (5)
SARL	**119**	105 (14)

Table 6
Survey results of all agents for the cleaning game.

Agent	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
ERL	**4.95**	**5.31**	**3.88**	4.43	4.69
PRL	5.75	5.45	3.4	4.09	5.22
SARL	5.37	5.62	3.76	4.68	4.93

5.2.3. The Revenge table

Finally, 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 agents the ERL agent (7.76), and the PRL agent (4.22); It can be seen that there is insignificant improvement then the DDQN agent. 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. 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. 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.

5.3. Results for the Stag-Hunt game

As for the Stag-Hunt game, we begin by comparing the average performance of each of the agents.

As shown in Table 7, SARL significantly outperforms all other agents ($p < 0.04$) in terms of its own utility. In addition, and similarly to the previous domains, SARL significantly outperforms all other agents ($p < 0.032$) in terms of social welfare. Unsurprisingly, humans interacting with the Follow-Stag baseline resulted in the highest performance. As for the ERL and the PRL agents, they both showed slight advancement over the DDQN agent. The ERL achieved a higher score in terms of agent's performance, but did not perform as well in terms of social welfare. The PRL was slightly less good compared to ERL at its own score, but achieved a much better social welfare score. This suggests that the PRL agent was more cooperative than the ERL agent, as can been seen in table 9.

Table 7
A comparison between the performance of each of the agents, along with the human player, who played against each of them, for the Stag-Hunt game.

Agent	Avg. agent's score	Avg. human's score	Avg. social welfare
Closest	27.08	23.98	51.05
Follow-Stag	26.29	**36.83**	63.12
Random	8.02	22.3	30.32
DDQN	27.85	29.19	57.04
ERL	30.63	24.33	54.96
PRL	28.25	32.75	61
SARL	**33.97**	35.31	**69.28**

5.3.1. The imperfection of the cleaning environment

In the following, we evaluate the prediction of the Stag-Hunt environment on all agents. Table 8 presents the predicted value compared to the actual score of every agent. As can be seen in the Table, once again the predicted value that is based on a human model indicates that the DDQN is the best policy, while, in practice, its performance it received a much lower score. In addition, Table 8 demonstrates that a human model that is based only on human interactions is not accurate enough, which necessitates the need for other solutions.

Table 8
The accuracy of the prediction of all policy agents in the Stag-Hunt environment.

Agent	True score	Prediction (error)
Closest	27.08	30.67 (3.59)
Follow-Stag	26.29	28.56 (2.27)
Random	8.02	7.73 (0.29)
DDQN	27.85	**33.67** (5.82)
ERL	30.63	30.41 (0.22)
PRL	28.25	33.35 (5.1)
SARL	**33.97**	33.37 (0.6)

5.3.2. The survey results

We now turn to analyze the survey results of the stag-hunt game as they appear in Table 9. As can be seen, the participants stated that SARL collaborated the most as compared to all agents. On the remaining features SARL obtained similar results compared to the DDQN agent. Moreover, the participants stated that the PRL agent collaborated more and acted more wisely than ERL agent. The β value for SARL was computed using Equation 5, and was set to 0.6 in this game. The intuition behind using Equation 5 in this game is that a value of $\beta < 0.5$ means that

Table 9
Survey results of all agents for the Stag-Hunt game.

Agent	selfish	comp.	coll.	wise	pred.
Closest	4.73	5.63	3.28	4.3	4.2
Follow-Stag	**2.83**	5.55	5.4	5.02	**4.9**
Random	4.24	5.82	3.39	3.08	3.57
DDQN	3.15	**4.81**	4.69	**5.12**	4.15
ERL	5.08	5.58	3.08	3.88	4.29
PRL	3.25	4.83	5.29	4.54	4.04
SARL	3.14	5.41	**5.48**	5.1	4.62

SARL would prefer the other agent to forage for fruits rather than itself, which is unreasonable. As for hunting the stag, the β value does not affect it, since both sides obtain the same reward.

6. Conclusions

In this paper, we present SARL, ERL and PRL. We showed that when data is limited, building an accurate human model is very challenging and a reinforcement learning agent, based on this data, does not perform well in practice. However, we showed that a social agent, i.e., an agent that tries to maximize a linear combination of the human's utility and its own utility, achieves a high score, and significantly outperformed other agents, including an agent that simply tries 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. Moreover, we proposed two additional methods that do not use the human reward function. The first method, the Empathy based RL (ERL) agent, builds a human model that also tries to learn the attitude of the human toward the agent. For this purpose, in addition to the actions of the participants, the ERL also requires to obtain answers to an opinion survey, presented at the end of the game, about the autonomous agent that the humans have interacted with. The second method, the Pessimistic RL (PRL) agent, tackles the overestimation bias by drawing multiple actions from the human model and assumes that the next action of the human player will be the least beneficial action for the agent (among the actions drawn from the model).

7. Future Work

In future work, we intend to show that SARL, ERL and PRL perform well also when considering other, possibly very different, settings. One option for such a setting is a setting with a continuous action space. Another direction for future work is to focus on situations in which the human reward function is not available apriori. While the ERL and PRL agents do not require access to a human reward function, the absence of such a function 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 [42] 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

[1] A. Azaria, ChatGPT Usage and Limitations, *OSF* (2022). doi:10.13140/RG.2.2.26616.11526.
[2] G. Skinner and T. Walmsley, Artificial intelligence and deep learning in video games a brief review, in: *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, IEEE, 2019, pp. 404–408.
[3] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot et al., Mastering the game of Go with deep neural networks and tree search, *nature* **529**(7587) (2016), 484–489.
[4] F.-h. Hsu, M.S. Campbell and A.J. Hoane Jr, Deep Blue system overview, in: *Proceedings of the 9th international conference on Supercomputing*, 1995, pp. 240–244.
[5] A. Tversky and D. Kahneman, The Framing of Decisions and the Psychology of Choice, *Science* **211**(4481) (1981), 453–458.
[6] G. Loewenstein, Willpower: A Decision-theorist's Perspective, *Law and Philosophy* **19** (2000), 51–76.
[7] D. Ariely, G. Loewenstein and D. Prelec, "Coherent arbitrariness": Stable demand curves without stable preferences, *The Quarterly Journal of Economics* **118**(1) (2003), 73–106.
[8] C.F. Camerer, *Behavioral Game Theory. Experiments in Strategic Interaction*, Princeton University Press, 2003, pp. 43–118, Chapter 2.
[9] N. Peled, Y.K. Gal and S. Kraus, A study of computational and human strategies in revelation games, in: *AAMAS*, 2011, pp. 345–352.

[10] A. Azaria, Z. Rabinovich, S. Kraus and C.V. Goldman, Strategic Information Disclosure to People with Multiple Alternatives, in: *AAAI*, 2011.

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

[12] A. Tversky and D. Kahneman, Advances in prospect theory: Cumulative representation of uncertainty, *Journal of Risk and uncertainty* **5**(4) (1992), 297–323.

[13] Y. Gal and A. Pfeffer, Modeling reciprocal behavior in human bilateral negotiation, in: *Proceedings of the National Conference on Artificial Intelligence*, Vol. 22, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007, p. 815.

[14] V.S. Subrahmanian, *Heterogeneous agent systems*, MIT press, 2000.

[15] A. Rosenfeld and S. Kraus, Using aspiration adaptation theory to improve learning, in: *AAMAS*, 2011, pp. 423–430.

[16] A. Azaria, Y. Aumann and S. Kraus, Automated strategies for determining rewards for human work, in: *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[17] J. Ackermann, V. Gabler, T. Osa and M. Sugiyama, Reducing overestimation bias in multi-agent domains using double centralized critics, *arXiv preprint arXiv:1910.01465* (2019).

[18] A. Azaria, Irrational, but Adaptive and Goal Oriented: Humans Interacting with Autonomous Agents, in: *IJCAI*, 2022.

[19] I. Shapira and A. Azaria, Autonomous Agents for The Single Track Road Problem, in: *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2021, pp. 81–85.

[20] M. Elhenawy, A. Elbery, A. Hassan and H. Rakha, An Intersection Game-Theory-Based Traffic Control Algorithm in a Connected Vehicle Environment (2015), 343–347. doi:10.1109/ITSC.2015.65.

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

[22] N. Jaques, A. Lazaridou, E. Hughes, Ç. Gülçehre, P.A. Ortega, D. Strouse, J.Z. Leibo and N. de Freitas, Intrinsic Social Motivation via Causal Influence in Multi-Agent RL, *CoRR* (2018).

[23] A. Peysakhovich and A. Lerer, Prosocial learning agents solve generalized Stag Hunts better than selfish ones, *CoRR* **abs/1709.02865** (2017).

[24] J.Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki and T. Graepel, Multi-agent reinforcement learning in sequential social dilemmas, *arXiv preprint arXiv:1702.03037* (2017).

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

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

[27] V. Vassiliades and C. Christodoulou, Multiagent reinforcement learning in the iterated prisoner's dilemma: fast cooperation through evolved payoffs, in: *The 2010 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2010, pp. 1–8.

[28] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J.Z. Leibo and N. De Freitas, Social influence as intrinsic motivation for multi-agent deep reinforcement learning, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 3040–3049.

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

[30] A. Azaria, Z. Rabinovich, S. Kraus, C. Goldman and Y. Gal, Strategic advice provision in repeated human-agent interactions, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26, 2012.

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

[32] J.R. Wright and K. Leyton-Brown, Beyond equilibrium: Predicting human behavior in normal-form games, in: *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[33] J.R. Wright and K. Leyton-Brown, Level-0 meta-models for predicting human behavior in games, in: *Proceedings of the fifteenth ACM conference on Economics and computation*, 2014, pp. 857–874.

[34] Y. Gal, A. Pfeffer, F. Marzo and B. Grosz, Learning Social Preferences in Games (2004).

[35] M. Cooper, J.K. Lee, J. Beck, J.D. Fishman, M. Gillett, Z. Papakipos, A. Zhang, J. Ramos, A. Shah and M.L. Littman, Stackelberg Punishment and Bully-Proofing Autonomous Vehicles, *CoRR* (2019).

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

[37] T.A. Carey, S.J. Tai and R. Griffiths, *Deconstructing Health Inequity: A Perceptual Control Theory Perspective*, Springer Nature, 2021.

[38] D. Ramachandran and E. Amir, Bayesian Inverse Reinforcement Learning., in: *IJCAI*, Vol. 7, 2007, pp. 2586–2591.

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

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

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

[42] A.Y. Ng, S.J. Russell et al., Algorithms for inverse reinforcement learning., in: *Icml*, Vol. 1, 2000, p. 2.

[43] E. Hughes, J.Z. Leibo, M. Phillips, K. Tuyls, E. Dueñez-Guzman, A. García Castañeda, I. Dunning, T. Zhu, K. McKee, R. Koster, H. Roff and T. Graepel, Inequity aversion improves cooperation in intertemporal social dilemmas, in: *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds, Curran Associates, Inc., 2018.

[44] R.S. Sutton and A.G. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[45] J.C. Rogan and H. Keselman, Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation, *American Educational Research Journal* **14**(4) (1977), 493–498.

[46] S.L. Zabell, The rule of succession, *Erkenntnis* **31**(2) (1989), 283–321.

[47] A. Azaria, Z. Rabinovich, S. Kraus and C.V. Goldman, Giving advice to people in path selection problems, in: *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[48] A. Rosenfeld, A. Azaria, S. Kraus, C.V. Goldman and O. Tsimhoni, Adaptive advice in automobile climate control systems, in: *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[49] T. Nguyen, R. Yang, A. Azaria, S. Kraus and M. Tambe, Analyzing the effectiveness of adversary modeling in security games, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27, 2013.

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

[51] A. Shvartzon, A. Azaria, S. Kraus, C.V. Goldman, J. Meyer and O. Tsimhoni, Personalized alert agent for optimal user performance, in: *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[52] A. Azaria, Z. Rabinovich, C.V. Goldman and S. Kraus, Strategic information disclosure to people with multiple alternatives, *ACM Transactions on Intelligent Systems and Technology (TIST)* **5**(4) (2015), 64.

[53] M. Bitan, Y. Gal, S. Kraus, E. Dokow and A. Azaria, Social rankings in human-computer committees, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27, 2013.

[54] A. Houenou, P. Bonnifait, V. Cherfaoui and W. Yao, Vehicle trajectory prediction based on motion model and maneuver recognition, in: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, November 3-7, 2013*, IEEE, 2013, pp. 4363–4369. doi:10.1109/IROS.2013.6696982.

[55] B. Kim, C.M. Kang, S. Lee, H. Chae, J. Kim, C.C. Chung and J.W. Choi, Probabilistic Vehicle Trajectory Prediction over Occupancy Grid Map via Recurrent Neural Network, *CoRR* **abs/1704.07049** (2017). http://arxiv.org/abs/1704.07049.

[56] N. Deo and M.M. Trivedi, Convolutional Social Pooling for Vehicle Trajectory Prediction, in: *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, IEEE Computer Society, 2018, pp. 1468–1476. doi:10.1109/CVPRW.2018.00196. http://openaccess.thecvf.com/content_cvpr_2018_workshops/w29/html/Deo_Convolutional_Social_Pooling_CVPR_2018_paper.html.

[57] C. Hermes, C. Wohler, K. Schenk and F. Kummert, Long-term vehicle motion prediction, in: *2009 IEEE Intelligent Vehicles Symposium*, 2009, pp. 652–657.

[58] W. Ding, J. Chen and S. Shen, Predicting Vehicle Behaviors Over An Extended Horizon Using Behavior Interaction Network, *CoRR* **abs/1903.00848** (2019). http://arxiv.org/abs/1903.00848.

[59] R. Chandra, U. Bhattacharya, A. Bera and D. Manocha, TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions, *CoRR* **abs/1812.04767** (2018). http://arxiv.org/abs/1812.04767.

[60] R. Chandra, T. Guan, S. Panuganti, T. Mittal, U. Bhattacharya, A. Bera and D. Manocha, Forecasting Trajectory and Behavior of Road-Agents Using Spectral Clustering in Graph-LSTMs, *IEEE Robotics Autom. Lett.* **5**(3) (2020), 4882–4890. doi:10.1109/LRA.2020.3004794.

[61] F. Leon and M. Gavrilescu, A Review of Tracking and Trajectory Prediction Methods for Autonomous Driving, *Mathematics* **9**(6) (2021), 660.

[62] N.K. Jong and P. Stone, Model-based function approximation in reinforcement learning, in: *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 2007, pp. 1–8.

[63] O. Amir, D.G. Rand and Y.K. Gal, Economic games on the internet: The effect of $1 stakes, *PloS one* **7**(2) (2012), e31461.

[64] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel and I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6382–6393.

[65] N. Bard, J.N. Foerster, S. Chandar, N. Burch, M. Lanctot, H.F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes et al., The hanabi challenge: A new frontier for ai research, *Artificial Intelligence* **280** (2020), 103216.

[66] D. Hadfield-Menell, S.J. Russell, P. Abbeel and A. Dragan, Cooperative inverse reinforcement learning, *Advances in neural information processing systems* **29** (2016), 3909–3917.

[67] L.P. Syll, Why game theory never will be anything but a footnote in the history of social science, *Real-World Economics Review* **83** (2018), 45–64.

[68] A. Azaria, Z. Rabinovich, S. Kraus, C.V. Goldman and O. Tsimhoni, Giving Advice to People in Path Selection Problems, in: *AAMAS*, 2012.

[69] K. Hindriks and D. Tykhonov, Opponent modelling in automated multi-issue negotiation using bayesian learning, in: *AAMAS*, 2008, pp. 331–338.

[70] A. Azaria, Z. Rabinovich, S. Kraus, C.V. Goldman and Y. Gal, Strategic Advice Provision in Repeated Human-Agent Interactions, in: *IJCAI*, 2012.

[71] I. Blanken, N. van de Ven and M. Zeelenberg, A meta-analytic review of moral licensing, *Personality and Social Psychology Bulletin* **41**(4) (2015), 540–558.

[72] F. Arabshahi, J. Lee, M. Gawarecki, K. Mazaitis, A. Azaria and T. Mitchell, Conversational Neuro-Symbolic Commonsense Reasoning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2021, pp. 4902–4911.

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

[74] I.T. Freire, X.D. Arsiwalla, J.-Y. Puigbò and P. Verschure, Modeling theory of mind in multi-agent games using adaptive feedback control, *arXiv preprint arXiv:1905.13225* (2019).