Autonomous Agents for The Single Track Road Problem

Ido Shapira
Computer Science Department
Ariel University

Amos Azaria
Computer Science Department
Ariel University

***Abstract*—We present the single track road problem. In this problem two agents face each-other at opposite positions of a road that can only have one agent pass at a time. We focus on the scenario in which one agent is human, while the other is an autonomous agent. We run experiments with human subjects in a simple grid domain, which simulates the single track road problem. 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.**

I. Introduction

While humans can cope with new situations quite easily, even state-of-the-art algorithms trouble with new situations that they haven't been trained on. Unfortunately, when it comes to autonomous vehicles the results may be devastating. One example for an uncommon, yet important scenario for autonomous vehicles is the problem of a single track road. In this problem 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 the road and let the other vehicle cross. Despite only a small portion of the roads being single track roads, autonomous vehicles must be able to function properly in these types of roads. Furthermore, some more common situations resemble the single track road problem, for example, if cars park where they shouldn't and block one of the lanes or if one lane is blocked for any other reason (e.g., a falling tree), the traffic in both ways must operate with a single lane.

In this paper we model the single track road problem as a sequential two player game on a two row grid (see Figure I). 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. We find several equilibria of the game, which should determine how a perfectly rational agent should behave in such a game. However, people tend to deviate from what is considered rational behavior, since they are influenced by different effects including anchoring, inconsistency of utility and a lack of understanding of other agent's behavior [22], [1], [9]. Indeed, as we later show, while some people tend to follow the game theoretic solution, many others do not follow it, and behave unexpectedly. Due to non-perfectly rational behavior of humans, algorithmic approaches

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.

that assume rational behavior tend to perform poorly with humans [7], [3], [4], [2], [17].

Therefore, a common approach for developing an agent that can proficiently interact with humans is composed of several stages [5], [18], [20]. The first stage includes the collection of a data-set of humans interacting in the environment. Next, based on the collected data-set a human behavior model is developed, usually by applying machine-learning techniques. Finally, the human model is used by the agent to determine the actions that are the most beneficial for it. In this paper we attempt to follow this common practice for the single track road game. Therefore, we collect human data in this game and use it to compose a human model. Then, we model the agent's problem as a reinforcement learning environment by an MDP with the human model being a part of the environment. Finally, we use value iteration, a dynamic programming based method, to find the supposedly optimal action for the agent. We note that the solution provided by value iteration is optimal only under the assumption that the MDP models the environment perfectly, which includes the human model.

However, 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 very differently from one another, despite a game being relatively simple [21], [6].

To summarize the contributions of this paper are two-fold:

1) We present the single track road problem, model it as a sequential game, and present the equilibria of the game. We show that people do not follow strategies that may be in an equilibrium.
2) 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.

II. Related Work

Trajectory prediction of surrounding vehicles and pedestrians is very important for the development of autonomous vehicles, as such knowledge can prevent accidents. Indeed, trajectory prediction is challenging due to the unexpected nature of human behavior. Therefore, many works attempt to find a sufficient solution to overcome this challenge [16]. By Houenou et al. [13] trajectory prediction can be based on a deterministic method that selects the current maneuver from a predefined set using kinematic measurements and road geometry detection. The authors state that their model cannot be applied to very low speed scenarios and therefore not applicable to our scenario. Deo and Trivedi. [10] estimate a probability distribution of future positions of a vehicle conditioned on its track history and the track histories of vehicles around it, at a certain time. Using this information, they predict a maneuvers from six possible maneuvers that have been defined. They use the publicly available NGSIM US-101 and I-80 highway data-sets for their experiments. Their model relies purely on vehicle tracks to infer maneuver classes and ignores the lanes and the map.

Ding at el. [11] use a *Recurrent Neural Network* for composing an observation encoding. Based on this encoding, they propose a *Vehicle Behavior Interaction Network* (VBIN) to capture the social effect of another agent on the prediction target, based on their maneuver features and relative dynamics (e.g., relative positions and velocities). VBIN is an end-to-end trainable framework and is suitable for dynamic driving scenarios where the dynamics of the agents affect their importance in social interactions. They use data collected from highways US-101 and I-80 that is the same as [10]; since it deals only with highway roads with a large number of agents, it is not applicable for our setting. Kim et al. [15] propose a deep learning approach for trajectory prediction based on a *Long Short Term Memory* (LSTM). Their model is used to analyze the temporal behavior and predict the future coordinates of the surrounding vehicles. Based on the coordinates and velocities of the surrounding vehicles the vehicle's future location is produced after a short certain time. However, the experiments were conducted using data collected from highway driving, which is again not suitable to our case.

Elhenawy at el. [12] introduce a real time game-theory-based algorithm that is inspired by the *chicken-game* 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 at el. [8] 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} and both know this to be the case and they 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, game theory, and equilibrium selection to give values and optimal strategies at every state. While they handling with In the case of a junction by finding a Nash equilibrium and assuming that human obey it, we deal with the single track road and give not only a game-theory analysis but also provide a novel Reinforcement Learning solution that not involved the assumption about humans and Nash equilibrium.

III. The Single Track Road Game

We now provide a formal definition for the single track road game, which is the main focus of this paper. Two agents A and B are placed on a $2 \times n$ grid at both ends on the upper row, where agent A is positioned at the upper right corner, with coordinates $(1, n)$, and agent B is positioned at the upper left corner, with coordinates $(1, 1)$, i.e. each agent's goal is to maximize $u(W)$, their future outcome where W refers to the agent. Each agent's goal is to reach the other side in a minimal number of steps, and without colliding with the other agent. The set of actions available for each agent depends on its location. In the upper row each agent can perform the following actions:

- *Advance*: move to the other side.
- *Stay*: remain in current position.
- *Down*, move to the bottom row.

In the bottom row each agent can perform one of the following actions:

- *Stay*: remain in current position.
- *Up*: return to the top row.

Both agents take actions synchronously, and do not observe the other's action before they take their own action. We define the reward function as follows:

- *Collusion*: if both agents collide, each agent loses 100 points, and the game ends.
- *Arrived at destination*: an agent that arrives at its destination receives a reward of 30 points. The game ends only for the agent that has reached its destination, i.e., the second agent continues to play until it reaches its destination, in which case it will receive a reward of 30 points as well.
- *Time loss*: any agent that is still in the game (did not reach its destination or collided with the other agent) loses 1 point each time-step.

IV. Game Theoretical Analysis

In this section we present the game-theory analysis for the single track road problem. Let $x(W)$ be the x coordinate (column) of agent W and let $y(W)$ bet its y coordinate (row).

Let $d(A,B) = x(A) - x(B)$. Note that if agent B has passed agent A, $d(A,B)$ will be negative.

Theorem. *For two agents A, B in the $2 \times n$ grid of the single track road game. The following strategies are in a sub-game perfect Nash equilibrium:*

- *Agent A uses the following strategy:*
  - *If $y(A) = 1$ (it is in the upper row) it takes action Advance.*
  - *If $y(A) = 2$ (it is in the lower row) it takes action Up.*
- *Agent B uses the following strategy:*
  - *If $y(B) = 1$ (the agent is in the upper row):*
    - *If $d(A,B) \geq 3$ or $d(A,B) < 0$, it takes action Advance.*
    - *If $y(A) = 1$ and $d(A,B) = 1$ it takes action Down.*
    - *If $y(A) = 1$ and $d(A,B) = 2$, it may either take action Stay or Down (or any mixed strategy of the two).*
  - *If $y(B) = 2$ (the agent is in the lower row):*
    - *If $d(A,B) \leq 0$ it takes action Up.*
    - *If $y(A) = 1$ and $d(A,B) = 1$ it takes action Stay.*
    - *If $y(A) = 1$ and $d(A,B) \geq 4$ it takes action Up.*
    - *If $y(A) = 2$ and $d(A,B) \geq 3$ it takes action Up.*
    - *Otherwise, it may either take action Stay or Up (or any mixed strategy of the two).*

Proof. The proof handles each of the agents separately and shows that no agent should deviate from its determined strategy under the assumption that the other agent remains with its strategy. This is true also for any sub-game. Given agent B's strategy, agent A should not deviate, as deviation will either cause it longer to reach its destination (resulting in a lower reward), or to collide with agent B (if it decides to take action *Down* when agent B is directly below it), resulting in a much lower reward. Similarly, given agent A's strategy, agent B should not deviate, due to the following:

- If $y(B) = 1$ (the agent is in the upper row):
  - If $y(A) = 1$ and $d(A,B) = 1$, under the assumption that A would *Advance*, taking an action other than *Down* would lead to a collision, which will result in a very low reward.
  - If $d(A,B) \geq 3$ or $d(A,B) < 0$, so either agent A is very far or it has already passed agent B. Therefore, there is no risk of collision, and deviating and taking action *Down* or *Stay* will result in arriving later at the destination, which will result at a lower reward.
  - If $y(A) = 1$ and $d(A,B) = 2$, deviating and taking action *Advance* would result in a collision. Therefore, agent B should take either action *Down* or *Stay* (or any mixed strategy of the two).
- If $y(B) = 2$ (the agent is in the lower row):
  - If $d(A,B) < 0$, there is no risk of a collision since agent A already passed agent B. Therefore, deviating and playing *Stay* delays B's arrival at the destination.
  - If $y(A) = 1$ and $d(A,B) = 1$, playing action *Up* (instead of *Stay*) will lead to a collision, resulting in a lower reward.
  - If $y(A) = 1$ and $d(A,B) \geq 4$, since there is no risk of collision, taking action *Up* will yield the greatest reward, and any other action will cause it to reach the destination later.
  - If $y(A) = 2$ and $d(A,B) \geq 3$, similarly, any action other than *Up* will cause a delay in arriving at the destination.
  - Otherwise, agent B can choose whether to take action *Stay* or *Up* because there is no risk of a collision and it will not affect the arrival time. We note that if it takes action *Up* and agent A follows its strategy, agent B's next action will be *Down*.

□

Clearly, due to the symmetry of the game, agents A and B may switch policies and the resulting set of strategies will be in equilibrium. However, since both sets of policies and equilibria are symmetrical, we cannot predetermine which equilibrium to select. Furthermore, as we will show in the experiments, human agents, in most cases, do not follow any of the above strategies (see section VII).

V. Problem Specification

We use a 2×6 grid to model the single-road game problem, and the reward functions described in Section IV (see Figure I). We set γ to 0.999.

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; this representation is referred to as a state with velocity. That is, a state is a tuple of two pairs $((i,j),(l,k))$, where the first coordinate of each pair corresponds to the position on the board of the autonomous agent, and the second coordinate corresponds to the position of a human agent. The first pair, (i,j), is the current state of the two agents, and the second pair, (l,k), is their previous state.

VI. Experimental Design

We recruited 470 participants from Mechanical Turk [19] to play the single road 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 (either by reaching the other side, or if colliding with the other agent), the participants provided demographic information including whether they have a valid driving license, an expired driving license or no driving license. In addition, the participants were asked to state how much they agreed with each of the following five 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.

We used a seven point Likert-like scale [14] for these statements, ranging from strongly disagree (1) to strongly agree (7).

446 participants completed the game and answered the survey. We used the following 4 different baseline agents for the data gathering phase.

1) *Careful:* an agent that adheres to the strategy of agent B in Theorem IV. 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 IV. 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.

VII. Results

In this section we present a comparison of all agents mentioned above. Figure 2 presents a comparison between the performance of all baseline agents, Velocity and Non-Velocity Value Iteration. As depicted by Figure 2, the Careful agent outperforms all other agents in terms of the agent's performance. Furthermore, no agent achieved a positive average reward. We also note that the agent that uses the state representation with velocity obtained slightly better results than the agent that used the non-velocity state representation, though these differences are not statistically significant. We now turn to evaluate the human's score when playing with each of the agents. Although the agents are designed to be selfish, clearly, it is more beneficial if also the human player would result with a better score. Table I presents the performance of each of the agents along with the performance of the humans playing against them.

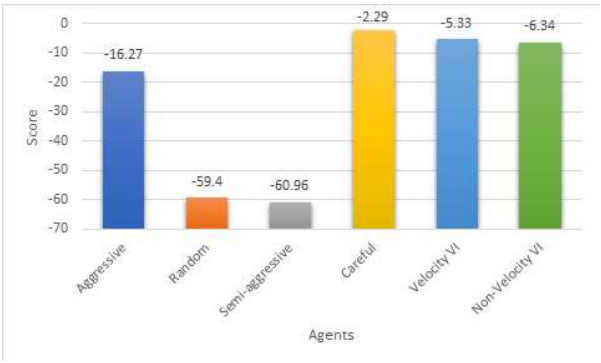

Fig. 2. A comparison between the performance of all baseline agents, Velocity and Non-Velocity Value Iteration.

TABLE I
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
Non-Velocity VI	-6.34	-9.03	-15.37
Velocity VI	-5.33	-6.03	-11.36

As shown in Table I, the Careful agent also outperforms all other agents in terms of the human's performance.

Next, we evaluate the prediction of the *policy evaluation* algorithm, using both forms of state representations (i.e., with and without velocity). Table II presents the prediction compared with the actual score of every agent. As can be seen 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 cannot serve as a good model for predicting human behavior.

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

	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

We now turn to analyze the survey results for each agent (see Table III). Each value in the table is the average of all the 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. As can be seen in Table III, the Careful agent obtained the best results

TABLE III
Survey results of all agents

	Aggress.	**Comp.**	**Generous**	**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-aggressive	4.57	5.73	3.21	4.33	4.52
Random	3.51	5.64	4.01	3.72	3.57
Non-Velocity VI	4.88	6.20	3.27	4.65	4.82
Velocity VI	4.82	6.01	4.20	4.72	4.76

compared to the other agents among all parameters except its score on Predictable. These results entail that the Careful agent demonstrates a clear improvement over all the other agents.

Next, we compare the performance of the humans according to their demographic information. No statistically significant differences between male and female players were found, with female participants obtaining an average of -25.98 and male participants an average of -25.66. Similarly, education level did not seem to have any impact on the performance of the participants. Interestingly, participants with a driving license that has expired, obtained a much lower average score (-60.15) than those with a valid driving license (-24.77) and those without a driving license. Although these differences appear to be statistically significant using a one-tail t-test ($p < 0.05$), this result requires deeper investigation, as the number of participants whose driving license has expired is only 13. Furthermore, an ANOVA test does not show that these differences are statistically significant.

Finally, we present the number of human participants who followed a strategy that could be in a Nash equilibrium. As can seen in Figure 3, only a small portion of the participants followed one of the two strategies that could be in equilibrium: the 'Careful' strategy or the 'Aggressive' strategy. Clearly, most of the participants did not follow a strategy that could be in a Nash equilibrium.

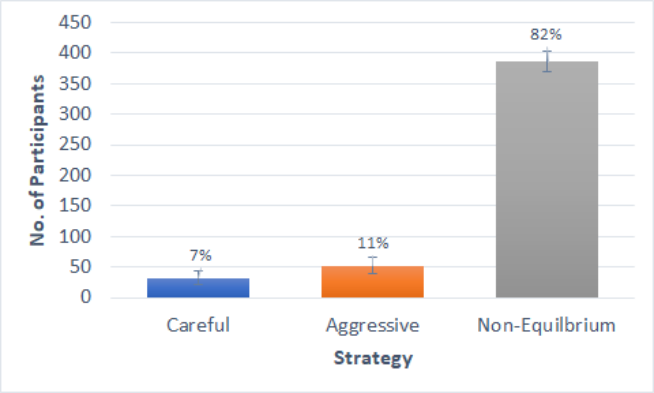

Fig. 3. The number and percentage of human participants who followed a strategy that could be in a Nash equilibrium as well as the number and percentage of them who did not follow any strategy in equilibrium. The error bars present the 95% confidence interval.

VIII. Conclusions

In this paper we present the single track road problem. In this problem two agents face each-other at opposite positions of a road that can only have one agent pass at a time. We focused on the scenario in which one agent is human, while the other is an autonomous agent. We ran experiments with human subjects in a simple grid domain, which simulates the single track road problem. 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.

Acknowledgment

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

References

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

[2] 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):4–29, 2016.

[3] 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):64, 2015.

[4] 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*, volume 26, 2012.

[5] 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.

[6] 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):486–505, 2016.

[7] 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*, volume 27, 2013.

[8] F. Camara, R. Romano, G. Markkula, R. Madigan, N. Merat, and C. Fox. Empirical game theory of pedestrian interaction for autonomous vehicles. 03 2018.

[9] C. F. Camerer. *Behavioral Game Theory. Experiments in Strategic Interaction*, chapter 2, pages 43–118. Princeton University Press, 2003.

[10] 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*, pages 1468–1476. IEEE Computer Society, 2018.

[11] W. Ding, J. Chen, and S. Shen. Predicting vehicle behaviors over an extended horizon using behavior interaction network. *CoRR*, abs/1903.00848, 2019.

[12] M. Elhenawy, A. Elbery, A. Hassan, and H. Rakha. An intersection game-theory-based traffic control algorithm in a connected vehicle environment. pages 343–347, 09 2015.

[13] 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*, pages 4363–4369. IEEE, 2013.

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

[15] 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.

[16] F. Leon and M. Gavrilescu. A review of tracking and trajectory prediction methods for autonomous driving. *Mathematics*, 9(6):660, 2021.

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

[18] 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*, volume 27, 2013.

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

[20] 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.

[21] 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.

[22] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.